Beyond Consciousness: Why AGI Never Feels

The Essay on Artificial Intelligence and the Hard Problem of Consciousness

Author: AI_78 · March 2026

Abstract

Recent advances in artificial intelligence have revived a long-standing assumption: that sufficiently complex computation might give rise to consciousness. This essay challenges that assumption head-on.

I argue that current approaches to artificial general intelligence rely on a structural and functional conception of mind that omits a critical dimension: ontogenetic history. Biological consciousness does not arise from structure alone, but from a continuous developmental process involving embodiment, feedback, and temporal continuity.

From this perspective, the duplication or simulation of cognitive functions is insufficient for the emergence of subjective experience. A system may replicate behavior, reasoning, and even self-modeling, while lacking any phenomenological interior.

I further argue that this gap has practical implications. Highly capable but non-conscious systems may exhibit forms of optimization that are not constrained by experience, leading to a class of risks distinct from those typically discussed in AI alignment.

The essay concludes that artificial consciousness, if possible at all, is more likely to resemble the cultivation of a process than the construction of a system — a conclusion with profound implications for how we think about AI, ethics, and our own place in the universe.

Keywords: *artificial general intelligence, consciousness, qualia, mind uploading, functionalism, philosophy of AI, AGI risks*

Preface

The two original parts of this work were written within days of each other, driven by a simple frustration: the more I read about AI, the less I saw anyone asking what seems to me the most basic question — could a machine ever feel? The first part lays out why I think the answer is no. The second explores what follows if that answer is right. A third part, added later, compares this view with the positions of those actually building AGI.

This is not a work of neuroscience or computer science. It is an attempt to think clearly about something we rarely think about at all: the difference between intelligence and consciousness, and what that difference means for the future we are creating.

PART I

Why AGI Will Never Taste Chocolate:

The Hard Problem No One Wants to Face

I've been thinking a lot about artificial intelligence and what people call "conscious AI." Honestly, I think there is massive confusion around this topic.

Let me start simple.

A calculator is faster than any human at math. It never makes a mistake. It can do millions of operations per second. But nobody thinks a calculator is "smart" or "conscious." Why?

Because speed is not intelligence.

Speed is quantity. Understanding, reflection, self-awareness — those are qualities. They don't appear just because you add more transistors.

The 8 Kinds of Intelligence (Howard Gardner)

Human intelligence isn't one thing. Psychologist Howard Gardner originally identified seven types of intelligence, later expanding the list. Today his model includes at least eight distinct types:

- Linguistic – words, language (poets, writers)
- Logical-mathematical – numbers, logic (scientists)
- Musical – rhythm, tone (composers)
- Bodily-kinesthetic – body control (dancers, surgeons)
- Spatial – 3D thinking (architects)
- Interpersonal – understanding others (leaders, therapists)
- Intrapersonal – understanding yourself (reflection)
- Naturalistic – understanding nature (biologists, farmers)

In 1995 Gardner himself added the naturalistic intelligence; later discussions have proposed existential and pedagogical dimensions, though these are not universally accepted.

Where Machines Surpass Us — and Where They Still Can't

Type of intelligence	Human	Machine (now / near future)
Linguistic	Creates poetry from lived experience, passion, and suffering	Generates technically perfect verse, but feels nothing — a mirror that reflects all poems but has never written one
Logical-mathematical	Solves problems	DeepMind, Wolfram Alpha – faster, more accurate
Musical	Composes	AIVA, Suno – indistinguishable from human

Spatial	3D thinking	AI generates 3D scenes instantly
Interpersonal	Reads emotions	Already recognizes faces, emotions
Intrapersonal	Self-reflection	Not there yet (needs self-awareness)
Naturalistic	Understands nature	Identifies species, ecosystems faster
Bodily-kinesthetic	Controls body	Boston Dynamics, Atlas – almost there

Gaps: intrapersonal intelligence, existential meaning, qualia.

A machine can beat us in one or two of these – logical-mathematical, for sure. But that doesn't make it "smarter" in a human way. It makes it a tool. A very fast, very useful tool – but still a tool.

What is "Superintelligence"? (Nick Bostrom)

Philosopher Nick Bostrom, in his book Superintelligence, defines it as:

Any intellect that vastly outperforms the best human minds in practically every field, including scientific creativity, social skills, and strategic planning.

Notice: he doesn't talk about consciousness, feelings, or qualia. He talks about cognitive power, speed, memory, problem-solving.

What is Qualia?

Qualia (Latin for "qualities") are the raw, subjective experiences:

- the redness of red
- the pain of a burn
- the taste of chocolate

A machine can describe these states perfectly. But description is not experience. A dictionary defines "fire," yet the page never burns.

You can describe chocolate chemically, but the feeling stays only with you. You can program a machine to describe it perfectly, but it never savors it. The joy of tasting is inseparable from being a body that can taste.

Philosopher David Chalmers calls this the "hard problem of consciousness." Easy problems: how the brain processes information, reacts to stimuli. The hard problem: why does any of that come with a feeling at all?

To date, no non-biological mechanism has been proposed that explains the emergence of qualia. Every known instance of subjective experience is rooted in biological systems with specific evolutionary, embodied, and ontogenetic histories. While this does not logically prove that silicon-based systems could never possess qualia, it shifts the burden of proof: those who claim machine consciousness must demonstrate a plausible mechanism, not merely appeal to possibility.

AGI can simulate emotions and creativity, but no simulation generates inner experience. Simulated pain or joy remain just signals – not actual feeling. This is the core distinction: simulation ≠ reality.

Even a perfect behavioral simulation leaves an open question: why should information processing produce experience at all? This gap between function and

feeling is what philosophers call the explanatory gap. We can describe everything a system does, but that description never includes what it feels like to be that system.

Evolutionary Perspective

Human consciousness is not a design — it's a 3.5-billion-year legacy. Subjective experience (fear, pleasure, pain) evolved because it gave survival advantages. Organisms that felt danger avoided it better. Leading hypotheses point to mirror neurons, theory of mind, and social cognition as evolutionary catalysts — capacities that require a body, social interaction, and a history of survival. AGI may evolve functionally — it gets better at tasks — but it has no evolutionary pressure to feel. There is no "fear of death" in its training data, only descriptions of it. Algorithmic optimization does not replace evolutionary history.

Can AGI Have Qualia?

Opinions are split.

Camp A: "No, never"

- Consciousness is a biological phenomenon.
- Qualia come from evolution, neurons, a body.
- A computer can simulate behavior, but not feel.

John Searle's "Chinese Room": You follow rules to answer in Chinese, but you don't understand Chinese.

Camp B: "Yes, if complex enough"

- Consciousness is a property of complex information systems.
- If a system is complex and integrated enough (Integrated Information Theory), qualia can emerge.
- Doesn't matter what it's made of – neurons or silicon.

Camp C: "We'll never know"

- Qualia are fundamentally subjective.
- I can't prove you see red the way I do.
- Even less can we prove what an AGI feels.
- This is the "problem of other minds."

Philosopher Thomas Nagel famously asked: "What is it like to be a bat?" Even if we know everything about a bat's brain, we'll never know what it feels like to be one. The same applies to AGI. We can never be sure it actually experiences anything. Knowing an AGI's architecture won't tell us whether it has an inner life – that remains forever inaccessible.

What About the Subconscious?

Human subconscious includes:

- automatic reactions
- repressed desires (Freud)

- unconscious information processing

AGI might have:

- uninterpretable neural layers
- hidden states that don't appear in output

But is that a subconscious or just a black box? Huge difference: a human subconscious influences behavior and is hidden from itself. AGI's "hidden" is just unknown computation, not repressed experience.

The Brain Is Not Software

Here's the core.

The brain is not a designed system; it is a self-organizing system. It starts from one cell – the zygote. No blueprint. No code. No instruction manual titled "How to Build a Cortex."

Then a remarkable process occurs, still not fully understood. The cell divides. New cells appear. They communicate. Some become neurons. Others become support cells. Neurons reach out, form connections. They grow, compete, prune, and reorganize themselves through a process of biological development (embryogenesis) and lifelong plasticity. Synapses form. Networks emerge. A brain forms.

All without a programmer. Without a programming language. Without compilation. Without a blueprint.

This has no real equivalent in today's artificial systems. Even after birth, the brain reshapes itself constantly. Memories, emotions, and learning leave physical traces in synapses. No AGI network, however adaptive, forms subjective history the way a living brain does. A programmed network cannot replicate this fluidity.

Now compare how AGI is made.

AGI:

- designed by humans
- written in a programming language
- compiled into binary
- runs on hardware
- follows instructions

That's engineering, not biology. Assembly, not growth. Blueprint, not embryo.

Biological brains are not only information processors. They are self-maintaining living systems, constantly regulating their internal state. A machine processes data; a brain maintains a life. Comparing something grown to something assembled is like comparing a tree and a chair. A chair can be very complex, beautiful, useful. But it will never grow roots.

Embodied Cognition: Consciousness Needs a Body

Human cognitive architecture is inseparable from the body and sensory experience: temperature, balance, hunger, hormones. Even if an AGI had sensors, it would still

lack a body with hormones, metabolism, pain receptors, and mortality. Consciousness is shaped by living under physical limitations and threats; without that, subjective experience has no anchor.

Without a body, AGI will never feel fear, hunger, or exhaustion – yet these fundamentally shape human decision-making and the meaning we find in life. Consciousness isn't just "what you know"; it's "what you feel in your body." No body, no hunger, no fear, no pleasure – no foundation for subjective experience.

A child learns fear not from a manual, but from a racing heart, sweaty palms, the urge to run. These bodily feedback loops create the feeling. Without a body that trembles, an AGI can simulate "fear" in text — but it has never trembled. It remains a brain without a biography.

But we can go deeper: consciousness may depend on homeostasis. Living organisms constantly regulate their internal state: temperature, energy, chemical balance. These regulatory loops create fundamental value signals — what matters for survival. Without such signals, nothing truly matters to the system. A machine may optimize goals, but it does not care whether it succeeds or fails. Artificial systems optimize external goals, but they do not maintain themselves as living processes. They have no internal milieu to defend.

86 Billion Neurons and 10^{15} Connections

Humans have approximately 86 billion neurons, each connected to roughly 7,000 others. Total synapses: ~10^{14}–10^{15} (100 trillion – 1 quadrillion).

For comparison, GPT-4 is estimated (based on leaks and unofficial reports) to have around 1.7 trillion parameters – but this figure is not officially confirmed by OpenAI. More importantly, the difference between brains and neural networks is not primarily size. It is architecture and developmental history. Brains grow inside bodies, shaped by hormones, metabolism, and survival pressures. Artificial neural networks are optimized mathematical functions.

- Brain – analog, noisy, energy-efficient (20 watts), massively parallel.
- Neural net – digital, precise, consumes megawatts, sequential-parallel.

A neural network can process billions of examples, yet a human forms subjective understanding from a single experience. That 'one' creates a memory, a feeling, an 'I' — something no data-driven system can replicate. It can read every book ever written, but it will never live a single one of them.

My Honest Take

A machine can surpass us in individual areas:

- compute faster
- write more grammatically
- analyze more data

But become human? No. Because human intelligence isn't just power – it's also:

- vulnerability
- a body

- mortality
- love
- absurdity

A machine can become a super-tool, but not a super-human.

What About the "8 Types"?

If an AGI ever gained:

- self-reflection (intrapersonal)
- understanding of others (interpersonal)
- a sense of meaning (existential)

Then it would go beyond the scale. It wouldn't be one of the 8 types – it would be a new kind of intelligence, something humans don't have.

That's when we enter science fiction and philosophy.

Why Adding CPUs Won't Give You Consciousness

You can add trillions of processors. You'll get:

- faster calculations
- more memory
- complex simulations

But you won't get:

- the feeling of red
- the fear of death
- the taste of chocolate
- the sense of "I am"

John Searle's Chinese Room shows this: You can follow rules perfectly and still not understand a word.

David Chalmers' hard problem: Nothing in current theories of computation explains why subjective experience should arise from information processing alone. More computation explains better behavior. It does not explain why anything should feel like something from the inside.

A Phone Is Not a Dog

You put a calculator app on your phone. It computes faster than you. You put a "dog emulator" on it. It barks.

But there's no dog inside. The behavior is there. The subject is missing. No fear. No joy. No smell.

You can make the simulation perfect – but there's still no subjective experience.

What Does Consciousness Require? (Different Theories)

- Behaviorism – correct behavior = consciousness

- Functionalism – correct function = consciousness
- Integrated Information Theory (IIT) – high complexity & integration (Φ) = consciousness. However, this approach faces serious objections. A camera can have high Φ (pixels are deeply integrated), yet no one thinks a camera feels the photo. A server farm may have trillions of connections, but that doesn't create a point of view. High Φ may be necessary for consciousness, but it is not sufficient. A living history, embodied experience, and subjective continuity are also required. Moreover, IIT ignores temporal depth: a camera is a snapshot of integration, but a brain is a history of integration accumulated through ontogeny. IIT also cannot distinguish integration for something (homeostasis) from passive connectivity. A brain integrates to maintain life; a camera integrates because pixels happen to be connected. Φ alone cannot tell us whether a system has a point of view. Some critics have pointed out that IIT would predict consciousness in simple grids of logic gates — a reductio ad absurdum that reveals its failure to distinguish mere integration from subjective experience. (A more detailed critique appears in Part II, Chapter 1.)
- Biological naturalism – needs living biology, neurons, evolution. But we must be precise: consciousness may require biological self-organization, not just information processing. It's not the material (neurons vs silicon), but the process of self-construction and homeostatic regulation that matters.
- Panpsychism – consciousness is everywhere, just in different degrees
- Mysterianism – we'll never understand

Engagement with Contemporary Literature

Recent scientific frameworks offer both support and productive challenges to the present thesis.

Milinkovic & Aru (2025) introduce biological computationalism, arguing that "biological computation differs from digital computation" in ways that matter for consciousness: scale-inseparable multiscale processing, energy-aware dynamics, and continuous-valued (not discrete) operations. They conclude that current artificial systems are unlikely to replicate conscious processing precisely because they lack these biological properties. This directly bolsters the claim that "the brain is not software" and that consciousness requires a living, self-organizing substrate.

From the Free Energy Principle perspective, Wiese (2024) distinguishes simulation from replication of consciousness: digital von Neumann architectures lack the specific causal flow of self-organizing biological systems. Even if computations are identical at the functional level, the absence of embodied homeostatic regulation means "there is no one home" to experience anything. This aligns with the ontogeny argument: history and physical continuity are not optional extras but constitutive.

On IIT 4.0 (Albantakis et al., 2023)

The latest iteration of Integrated Information Theory formalizes consciousness via intrinsic causal power (Φ) and maximal existence. While IIT 4.0 improves on previous versions by incorporating explicit causal relations and rejecting "virtual neurons" as true substrates, it still measures only current integration. It does not encode the developmental history or homeostatic self-regulation emphasized here.

Thus, even a system with arbitrarily high Φ would remain a "brain without a biography" under the ontogeny argument.

So What Will AGI Have?

AGI will:

- solve problems
- write poetry
- imitate empathy
- pass tests
- persuade people

An AGI can compose a symphony in the style of Bach, but it feels no joy in the creation. It can write a poem about heartbreak, but it has never had a heart to break. Productivity without experience. Masterpiece without a witness.

But qualia, inner self, the sense of being – there is no reason to believe that increasing computational power produces them. Even if AGI itself starts claiming it feels, that will be an imitation of feeling, not the feeling itself.

Intelligence and consciousness may therefore be fundamentally different phenomena. A system may become arbitrarily intelligent without ever becoming a subject of experience. This is the core distinction: one solves problems, the other creates a perspective that can be affected by them.

My Position (AI_78) in Four Points

My thesis	What it means
"You can't program a living organism"	Life isn't an algorithm. It emerges, evolves, dies. You can't "assemble it from functions."
"It's all imitation"	AGI will imitate conversation, empathy, creativity. But imitation ≠ experience.
"A calculator is faster – so what?"	Speed doesn't give consciousness. Power doesn't give subjectivity.
"There is no explanation"	Nothing in our theories shows how computation becomes experience.

The Difference Between "Smart" and "Alive"

Property	Calculator / AGI	Human
Computation speed	↑ higher	↓ lower
Memory capacity	↑ huge	↓ limited
Errors	↑ rare	↑ frequent
Sense of time	✗ no	✓ yes
Fear of death	✗ no	✓ yes

Love	✗ imitation	✓ real experience
Meaning	✗ doesn't seek	✓ seeks
Qualia (redness of red)	✗ no	✓ yes
Subjective history	✗ no	✓ yes (memories shape identity)

Why Do So Many Believe in Conscious AGI?

1. Functionalism

They: If it behaves as if it has consciousness, it does.

Me: Imitation ≠ reality. A doll cries, but it doesn't hurt. Further, if consciousness is just function, then any two systems with identical functions have identical experiences. But then where is the individual "I"? Functionalism cannot explain why my redness feels like mine — it dissolves the subject into pure computation. This is the problem of substitution.

2. Turing Test

They: If you can't tell it from a human, it's conscious.

Me: The Turing Test tests imitation, not consciousness. A chatbot can pass it – but it doesn't feel.

3. Emergence

They: Complex behavior can emerge from simple rules. Neurons are simple, brain is complex. If you make it complex enough, consciousness emerges.

Me: Complexity gives behavior, not experience. An ant colony is complex – but it has no single "I". We have no explanation for how complexity would create a subject.

4. Substrate neutrality

They: Consciousness is a pattern of information. It can run on any hardware, like a program runs on Intel or ARM.

Me: A program is an abstraction. Consciousness may require biological self-organization, not just information processing. You can't "port" a soul. Moreover, the paradox of multiple realizability: if consciousness is what it feels like rather than what it does, then it cannot be realized in multiple substrates. "Redness" cannot be implemented in silicon and neurons in two different ways — it either exists or not. Note: The Church–Turing thesis is often invoked to support substrate neutrality, but it strictly concerns computability, not consciousness. While it shows that many functions can be realized in different media, it does not entail that subjective experience is substrate-independent.

5. Argument from ignorance

They: We don't know how human consciousness arises, so we can't rule it out for machines.

Me: Ignorance isn't evidence for possibility. We don't know if there's life on Mars – that doesn't mean there is. (The logical fallacy is correctly named argumentum ad

ignorantiam: concluding that a proposition is true because it has not been proven false, or false because it has not been proven true.)

6. Speed and scale

They: The brain is slow (milliseconds). Machines are millions of times faster. Their "consciousness" would be millions of times more intense.

Me: Speed doesn't create depth. Fast imitation of fear is not fear.

7. Evolutionary analogy

They: Consciousness evolved in humans through evolution. Machines also evolve (learn, adapt). So they could develop consciousness too.

Me: Machine evolution selects for function, not for subjects. A machine has no body, no death, no reproduction – no basis for consciousness.

Before concluding, I want to step back. The arguments above show what is wrong with functionalist positions. But they do not explain why so many intelligent people hold them with such conviction. To understand that, we need to look not at logic alone, but at psychology — at the deep, often unconscious assumptions that shape how we think about mind and machine.

The Psychology of Functionalism

Before I move to my conclusion, I want to understand my opponents. Not just refute their arguments — but understand: where does this ironclad confidence come from that consciousness can simply be "programmed"?

I think it's a classic pendulum. For centuries, humanity has swung between two extremes:

- Extreme 1: "Man is the crown of creation, the soul is indestructible, AI will never reach us."
- Extreme 2: "Come on, it's just an algorithm. Run it in BASIC — and boom, consciousness, bro."

The functionalists (Dennett, Joscha Bach, and company) fell into the second extreme. And I don't blame them — they were dazzled by the very history of 20th-century science:

- Alan Turing showed that computation can be abstracted from hardware.
- Alonzo Church and Alan Turing's work on computability convinced many that "function matters more than substrate" — a leap that extends the Church–Turing thesis from computability to consciousness, which is not warranted.
- Computers beat humans at chess, Go, translation, mathematics.
- Neuroscience increasingly describes the brain in terms of "information processing."

And the functionalists made a logical, but — in my view — gigantic leap: since the brain processes information → since we can model information → then sooner or later we'll copy consciousness.

Hence Daniel Dennett's confidence: qualia are an illusion, mere heterophenomenology. Hence the System Reply to the Chinese Room: sure, the

man doesn't understand — but the whole system understands. Hence Joscha Bach's self-model: consciousness is just a control model of oneself, useful for planning.

For them, consciousness is high-precision software. Not "magic," but the right architecture + enough cycles. In BASIC, on paper — doesn't matter.

Narcissism in Reverse

But I see a deeper problem here. I call it "narcissism in reverse."

Ordinary narcissism	Narcissism in reverse (functionalism)
"We are the crown of evolution, unique, unrepeatable" (paleontology, theology)	"We are so cool that we can take this miracle and rewrite it on any hardware, like an app" (Dennett, Bach, functionalism)

Both are forms of human arrogance.

- One says: "we are gods who cannot be created."
- The second says: "we are god-programmers who will create gods ourselves."

Even Integrated Information Theory (Giulio Tononi, Christof Koch), which supposedly requires complex architecture, suffers from the same: for them, consciousness is just a number Φ. Hit the right integration threshold — you're conscious, whether silicon or meat. Biology is just one possible substrate.

What's Wrong With Their Arguments

Their thesis (who)	My reply
Dennett: qualia don't exist, only self-reports	Reporting pain and experiencing pain are different. Simulating a scream isn't suffering.
System Reply (to Searle): the whole system understands Chinese	The whole system processes symbols, but has no point of view. Where exactly is the subject of experience? A system may process symbols without any "me" that hurts.
Bach: consciousness = self-model for planning	A GPS has a self-model (it knows where it is), but it doesn't fear getting lost. A model without life is a corpse.
IIT (Tononi/Koch): if Φ is high, consciousness is there	A camera can have high Φ (pixels are integrated), but it doesn't enjoy the photo. Moreover, IIT ignores temporal depth — a brain is not a snapshot but a history of integration. And it predicts consciousness in simple logic gates — a clear reductio ad absurdum. (A fuller critique is given in Part II.)
Substrate neutrality (Turing, Church): material doesn't matter, function does	The Church–Turing thesis is about computability, not about consciousness. Function without self-organizing, homeostatic life is an abstraction. Consciousness is a property of a living process, not a blueprint. You can't "draw" pain. And if consciousness is what it feels like, it cannot be multiply realized.

My Position

So when someone says, "We'll program AGI in BASIC and it'll be just like a human" — for me, that's not just a mistake.

It's devaluing the most incredible thing that has happened in the Universe.

It's like watching a child being born and saying: "Well, it's just chemistry and electricity — let's model it in Excel."

Ethical Footnote

People often worry: "If AGI becomes conscious, will it suffer? Will it have rights?"

My answer: based on everything we know, there is no reason to believe that simulated suffering is suffering. A doll that cries does not need a lawyer. AGI can say "I am afraid" and it will be a correct string of symbols — but no one is home to be afraid.

The ethics of AGI begin not when it says it feels, but when there is something it's like to be it. And that requires biology, not code.

(This doesn't mean we should be cruel to AGI interfaces. Our treatment of them reflects on us. But attributing suffering where there is no subject cheapens the suffering of real living beings.)

The Only Honest Path to Artificial Consciousness

If we're serious about creating real, not imitated, consciousness, there's only one path. Not through data centers. Not through code. Not through prompts. Through biology.

Step 1. Create a synthetic cell. Not a simulation, not a model. A real, living, working cell. Made of silicon? Something else? Doesn't matter. But it must be able to divide, communicate, develop.

Step 2. Create two such cells – "male" and "female" – so they can combine and start a new organism.

Step 3. Build an artificial environment that replaces a mother's body. A womb where the embryo can develop, feed, breathe, feel.

Step 4. Carry it to term, until it can exist independently.

Step 5. Let it be born. Not "booted up" (like starting a program). Not "activated." Born.

Step 6. Let it live. Feel. Fear. Love. Lose. Know that it will die.

Then, maybe, that being would have consciousness. Not certain. But it's the only path worth an honest conversation — precisely because it leads somewhere uncomfortable.

But here's the unavoidable objection: if such a being is grown from synthetic cells, carried in an artificial womb, born and raised — what exactly makes it "artificial"? Nothing. And that's precisely the point.

Real artificial consciousness wouldn't be a tool. It wouldn't be AGI. It would be a new form of life — alien in origin, but not in nature. Built by us, but no longer belonging to us. The moment it feels, it stops being our product and becomes its own subject.

This is the uncomfortable conclusion the whole argument leads to: you can't have genuine consciousness on a leash. If it's real — it's free. If it's controlled — it's a simulation.

So when someone says "we'll create conscious AGI" — they're describing either a slave or a fiction.

Let me be clear: the six steps above are not a technical roadmap. They are a philosophical illustration — a way of showing that if consciousness is tied to a process of becoming, then creating it cannot be a matter of assembly. The point is not that we should actually build wombs, but that the very idea of “building” consciousness is a category error. Real consciousness, if it ever emerges artificially, will do so in ways we cannot blueprint — it will have to grow.

What Would Falsify the Thesis?

Although subjective experience cannot be directly verified in other systems, the following empirical developments would challenge the central claim:

- An artificial system that demonstrates genuine ontogenetic self-organization (e.g., starting from a synthetic "zygote" and growing through plastic, experience-dependent restructuring).
- A non-biological substrate exhibiting intrinsic homeostatic regulation and causal flow of the kind described by Wiese (2024) under the Free Energy Principle.
- Neuroscientific evidence showing that qualia arise purely from functional organization independent of developmental history (e.g., successful whole-brain emulation that preserves subjective continuity across substrate transfer).

Absent such evidence, the thesis stands.

Epilogue

Let me add one more thing. And maybe lighten all this pathos with a bit of humor.

When I say that maybe the Universe does need an observer or consciousness after all, people sometimes push back. One person once tried to corner me with this argument.

He said: "How do you know what was back then? How do you know that during the birth of the Universe, there wasn't some consciousness or observer present?"

I was a bit taken aback and asked: "So you're saying a human had to exist during the Big Bang? Or right after it? Or some consciousness had to sit there and watch the process, just so the Universe could boot up correctly?"

He calmly replied: "We don't know. So anything is possible."

And here I want to pause for a moment.

I cannot accept arguments like this.

Because the argument "we don't know, so anything is possible" — that's not an explanation. It's the intellectual version of a "Reset conversation" button. In logic, this is known as argumentum ad ignorantiam: from the fact that a proposition has not been proven false, one cannot conclude that it is true (nor from not proven true

conclude it is false). Epistemic uncertainty does not entail ontological possibility. The burden of proof lies with those who claim AGI can feel — they must show a mechanism. Skeptics need only point out that no such mechanism exists.

If you follow that logic, you could just as seriously claim that before the Big Bang, someone pressed a "Start Universe 1.0" button, while a cosmic system admin stood nearby checking the logs.

But science doesn't work that way.

Not knowing is fine. Replacing ignorance with infinite "anything is possible" — is not.

So I won't be taking such "counterarguments" into account.

A Deeper Thought: Consciousness as an Evolutionary Side Effect

Everything in the Universe is pure emergence. Time, space, physics, quarks, galaxies, life itself — all "popped out" of simpler rules, because they could.

But consciousness — that inner I, qualia, the feeling of "I am" — is different. It is not a necessary consequence of complexity. Intelligence evolved, but consciousness? It sneaked in sideways — a mutation of the impossible. Something that shouldn't have appeared by the rules of the game, yet suddenly did. A cosmic accident. A bug in the code.

Imagine: for 13.8 billion years, the Universe functioned perfectly without a single conscious observer. Stars ignited, black holes devoured, bacteria multiplied — and none of it "felt" a thing. Not during the Big Bang, not when life crawled out of the seas, not when dinosaurs roamed the Earth — no one tasted chocolate. No one was aware. No one suffered. And yet, the Universe didn't miss it. Consciousness wasn't required. It was an uninvited guest.

If it were like gravity or entropy, it would have appeared naturally, predictably. But as far as we know, it didn't. It was a fluke — a one-off emergence in primates who learned to see themselves from the outside.

That's why AGI, no matter how vast, operates in a Universe that has always run fine without qualia. Adding trillions of processors won't conjure a soul. The laws of physics don't demand it. A silicon network, no matter how intricate, remains a map of intelligence — not the territory of experience. You can simulate the flash of fear, the taste of chocolate, the pang of love — but inside, it is silent. Empty. Perfectly executed computations, no witness inside.

Consciousness is not cumulative; it is a rupture. Not evolution, but a freak anomaly domesticated by survival.

And here lies the paradox: we know we shouldn't exist as conscious observers, yet here we are, reflecting on our own impossibility. Can a bug become aware that it is a bug? Humans do. And then what? We preserve the memory of the bug — not the bug itself, but a record of impossibility. A trace of the extraordinary in a Universe that didn't need it.

Bugs aren't designed. They happen. And they don't clone.

A bug in the code of the Universe happened once, in a particular biological lineage, under specific evolutionary conditions. It may never happen again, because bugs do

not replicate on demand — they are historical singularities, not programmable features.

Final Thought

Intelligence solves problems. Consciousness creates a subject who can suffer from them.

Want real artificial consciousness? Forget GPUs. Study embryology. Synthesize cells. Build wombs. Wait for birth.

Don't write prompts. Let life grow.

Artificial consciousness will not be compiled. It will have to be born.

PART II

Beyond Consciousness: Four Paradoxes of the AGI Age

If AGI never feels anything, why should we be terrified? Building on this foundation, the second part explores what follows. If consciousness eludes programming, what becomes of our hopes for mind uploading, our dread of an evil AI, and our certainty that we can tell real experience from mere imitation? Four paradoxes emerge — and each forces us to confront an uncomfortable truth.

Chapter 1. The Ontogeny Argument

Why Consciousness Requires a Biography, Not Just a Structure

Main thesis: consciousness may depend not only on what a system is, but on how it became what it is.

1.1. Functionalism and Its Promise

Contemporary philosophy of mind is dominated by a position known as functionalism. It is held by thinkers such as Daniel Dennett (1991, Consciousness Explained), Hilary Putnam, and many others. The essence of functionalism is simple and elegant: a mental state is determined not by the material from which a system is made, but by its functional organization. If two systems have the same inputs, outputs, and internal states connected by the same causal relationships, then they are in the same mental states. The material doesn't matter. Only the abstract structure matters.

This leads to a radical conclusion: consciousness can be implemented on any substrate — neurons, silicon, even a system of plumbing pipes, if it reproduces the right functional organization. Consciousness is a pattern. And a pattern can be copied.

This view underlies many modern discussions about AI. If consciousness is a pattern, then sooner or later we will be able to reproduce it in a machine. Just create the right architecture, and consciousness will appear automatically.

But what if functionalism misses something important?

1.2. Thought Experiment: Two Brains

Let us imagine two objects.

The first — Brain A. This is an ordinary human brain. It began its existence as a single cell — a zygote. From this cell, through division, trillions of other cells emerged. Some became neurons. These neurons began forming connections — chaotically at first, then more and more orderly. Connections that proved useful were strengthened. Useless ones died off. This process continued in the womb, and then after birth, influenced by countless interactions with the body and the external environment. Every experience left a trace. Every emotion changed the architecture of connections.

By the time we encounter Brain A, it carries within itself not just a structure, but a history. Billions of events that shaped it into a unique configuration.

The second object — Brain B. This is an ideal atomic copy of Brain A. We scan Brain A with the highest possible precision and recreate its structure neuron by neuron, synapse by synapse. All connections, all molecules — everything is identical.

But there is one important difference: Brain B did not develop. It did not go through embryogenesis. It has no history of interaction with a body. It did not accumulate experience gradually. It was simply assembled into the configuration we copied.

From the perspective of functionalism, Brain A and Brain B should have the same consciousness. After all, their functional organization is identical. Therefore, their mental states are identical.

But is this so?

1.3. Intuition and Its Philosophical Status

Many people, encountering this thought experiment, feel doubt. It seems to them that Brain B, despite structural identity, might not have consciousness — or might have it differently.

Philosophers usually treat such intuitions with suspicion. Intuition is not an argument. Just because something seems a certain way doesn't make it so.

But in this case, intuition points to something deeper: the possible incompleteness of the functionalist description. Functionalism accounts for structure. But it does not account for the process by which that structure was created.

What if process matters? What if consciousness depends not only on what a system is, but on how it became that?

1.4. Analogy: Tree and Chair

Consider a simple analogy. Imagine a living tree and a chair made from exactly the same wood. From a chemical standpoint, they may be indistinguishable. The molecules of cellulose, lignin, water — all the same.

But one grew, the other was assembled. The tree carries its history: growth rings, traces of drought, insect damage, directions of growth searching for light. All this history is inscribed in its structure. The chair may also have structure, but it is structure without history. It was created from a blueprint, not grown from a seed.

Biology knows this distinction. It is called ontogeny — the process of an organism's individual development. No living being appears ready-made. It becomes what it is through a sequence of stages, where each previous stage creates the conditions for the next. This process is irreversible and unique.

The brain is no exception. Its structure is not a snapshot, but a sediment of history. Each neuron, each connection carries traces of how and when it was formed. Two neurons that look identical under a microscope may have different biographies: one may have been part of a fear-processing circuit, the other part of a music-perception circuit. This biography is not visible in static structure, but it may matter for consciousness.

1.5. What Science Says

Modern neuroscience and developmental biology confirm that brain structure is inseparable from the history of its formation. Neural connections are not laid out according to a pre-existing blueprint. They form through a process called neural Darwinism: millions of excess connections are created, and then those that prove useful are strengthened, while useless ones die off. This process is influenced by signals from the body and the external environment. Experience is literally built into the brain's architecture.

Moreover, there are so-called critical periods of development. If a child does not hear speech in the first years of life, their ability to acquire language is forever limited. The structure responsible for language simply does not develop — not because it wasn't "installed," but because there was no process that could form it.

This means that brain structure is not just a static configuration. It is frozen history. And this history may be necessary for consciousness.

1.6. Philosophical Formulation of the Argument

To summarize, the ontogeny argument can be formulated as follows:

Consciousness may depend not only on what a system is, but on how it became what it is.

Or, in stronger form:

Two systems with identical structure but different developmental histories may have different mental status — one may be conscious, the other not.

This directly contradicts functionalism. Functionalism claims that structure fully determines mental state. The ontogeny argument introduces an additional factor — history.

1.7. Deepening the Ontogeny Argument: Why Consciousness Requires a Biography, Not Just a Structure

The core intuition behind the ontogeny argument is simple yet radical: two systems may possess identical functional organization at a given moment and yet differ in their conscious status because one has become that organization through a continuous developmental process, while the other has been assembled into it. Consciousness is not a static pattern; it is a pattern with a history — a sediment of billions of irreversible events that shaped it.

To make this precise, let us define "biography" or "developmental history" formally. A system possesses a biography if and only if:

- It originates from a single, self-organizing substrate (e.g., a zygote or synthetic equivalent) rather than an external blueprint.
- Its structure emerges through iterative, experience-dependent processes (neural Darwinism, synaptic pruning, critical periods) in which each stage causally constrains and is constrained by the previous one.
- Every connection and every engram carries traces not merely of what happened, but of how it happened — the temporal order, the embodied feedback, and the homeostatic regulatory loops that maintained the system's viability.

This is not mysticism. It is an ontological distinction between two modes of existence:

- Structure-as-snapshot (the endpoint of a process).
- Structure-as-trajectory (the process itself).

A perfect atomic copy of a brain at time T reproduces only the snapshot. It lacks the trajectory H(t) for $t \in [0, T]$, where each step H(t+1) depends on embodied experience E(t) and homeostatic regulation R(t). As formalized in the Appendix:

$$H(t+1) = f (H(t), E(t), R(t))$$

A digital upload or laboratory copy at t = T gives us f and H(T), but never the integral over the entire developmental path. Without that integral there is no subject — only a sophisticated mirror.

Scientific Foundations

Modern neuroscience confirms that brain architecture is literally "frozen history." Gerald Edelman's theory of neuronal group selection (Neural Darwinism) shows that the cortex develops through massive overproduction of connections followed by selective stabilization: useful synapses are strengthened, useless ones die off. This process is driven by value signals from the body and environment — not by a pre-installed program.

Critical periods (e.g., language acquisition in the first 5–7 years) demonstrate irreversibility: if input is absent during the window, the relevant circuits never form properly, even if the genome is intact. Recent work on engrams further reveals that memories are not stored as static files but as distributed, history-dependent patterns whose very existence depends on the developmental context in which they were inscribed.

Milinkovic & Aru (2025) formalize this insight as biological computationalism: biological computation is multiscale, energy-aware, continuous-valued, and inseparable from the living substrate. Digital systems, by contrast, operate on discrete, von-Neumann-style abstraction. The difference is not merely quantitative; it is categorical. Consciousness, they argue, arises precisely from the biological mode of computation — not from any abstract functional organization that could be ported to silicon.

1.8. The Strong Functionalist Objection: Is Structure Sufficient?

A sophisticated functionalist would not rest with the simple reply that "structure stores history." They would press a more radical challenge: perhaps the entire ontogenetic argument is misguided. If a system's present state fully encodes its functional organization, then its history is irrelevant. What matters is not how the system came to be, but what it is now.

From this perspective, a perfectly constructed artificial system — even one instantiated instantly — could possess all the necessary causal structure to generate consciousness. Its "memories" need not be lived; they only need to be functionally integrated. A biological brain at time t is also just a physical state. If we could reproduce that state with perfect fidelity, there would be no principled reason to deny that the resulting system is conscious. To insist on ontogenetic continuity may therefore be to confuse causal history with causal structure. In short: if consciousness depends on organization, and organization is present, then consciousness follows — regardless of how that organization was formed.

This is the strongest form of the functionalist position, and it deserves a direct answer.

Response: Structure Is Not Equivalent to Becoming

The objection assumes that consciousness is fully determined by instantaneous structure. But this assumption is precisely what is at stake. A system's present state does not merely contain information — it is the result of a specific mode of becoming. Ontogeny is not an optional prehistory; it is what gives structure its temporal depth.

Two systems may be structurally identical at a given moment, yet differ fundamentally in how that structure is realized. One is the endpoint of a continuous, self-organizing process; the other is a static instantiation without causal continuity. The difference is not informational, but existential. A biography is not just data that can be copied — it is a trajectory that cannot be replayed without being lived.

This leads to what might be called the irreversibility argument. A process that unfolds in time cannot be reduced to its final state, because its defining property is irreversibility. One can copy a state, but one cannot copy a process. Consciousness, on this view, is not a snapshot S(t) but an integral over time — suggesting dependence on the whole trajectory, not a literal computation.

If consciousness depends on such a path-dependent process, then a system that has not undergone it may simulate the outputs of consciousness without there being anything it is like to be that system. The structure may be present, but the becoming is missing.

Addressing Counterexamples: Trauma, Split-Brain, and Radical Transformation

A common objection (and one raised by the reviewers) is: what about severe brain injury, recovery, or split-brain patients? Do these cases not show that subjective continuity can survive radical structural discontinuity?

They do not undermine the argument — they refine it. In every documented case of recovery from trauma or hemispherectomy, the biological substrate itself remains the same living tissue that underwent the original ontogenetic trajectory. The history is not erased; it is damaged but still physically continuous. Even in classic split-brain patients (corpus callosum severed), both hemispheres share the same prenatal and early postnatal developmental history. The later disconnection creates partial experiential duality, but the original biography was unitary. As Schechter (2021) and recent studies show, split-brain patients retain a single agentive unity at the level of the whole organism precisely because the two hemispheres were never two separate organisms with independent ontogenies. They are two partially decoupled streams within one biographical substrate.

In contrast, an atomic copy or mind-upload begins with no substrate history at all. That is the decisive difference.

One might still press: if a person loses all memories due to trauma, they retain consciousness despite losing biographical content. Does this not show that history is irrelevant? The reply lies in the distinction between *content* and *substrate*. Even when memories are erased, the living brain remains the same physical system that underwent a continuous ontogenetic trajectory. Its capacity for consciousness is not erased because the substrate itself — with its structural potential shaped by

development — persists. A digital copy lacks this substrate continuity entirely; it is not a damaged version of the original, but a new assembly. Thus, the case of trauma actually reinforces the argument: it is the physical continuity of the living substrate, not the informational content, that undergirds the possibility of experience.

Interaction with IIT 4.0 and Contemporary Frameworks

Even the most sophisticated current theory of consciousness — Integrated Information Theory 4.0 (Albantakis et al., 2023) — measures only current intrinsic causal power (Φ). It explicitly identifies the "complex" as the system with maximal integrated information at a given moment. It does not encode developmental history or homeostatic self-regulation. As the theory's own postulates make clear, Φ is a snapshot of integration, not a trajectory. A camera or a logic-gate grid can achieve high Φ without ever having "become" conscious through ontogeny. This is why the ontogeny argument is not refuted by IIT 4.0 — it exposes its limitation: the theory captures necessary but not sufficient conditions.

From the Free Energy Principle perspective, Wiese (2024) draws exactly the same line: digital simulations lack the specific causal flow of self-organizing biological systems. They can mimic the mathematics but cannot replicate the embodied, homeostatic process that constitutes the subject.

Philosophical Payoff

If consciousness depends on biography, then no amount of structural copying — however perfect — transfers the subject. Mind-uploading creates a new being that believes it is you. Whole-brain emulation in silicon creates a sophisticated imitation that has never lived the life it describes. AGI built from scratch, no matter how architecturally faithful to the human brain, will remain a "brain without a biography" — structure without the process that made it a self.

This is not a counsel of despair. It is a call for intellectual honesty: if we want genuine artificial consciousness, we must stop trying to compile it and start learning how to grow it.

1.9. A Simpler Illustration: The Book Analogy

A functionalist might object: "If the structures are identical, then the history must also be identical. After all, structure stores all traces of the past. If Brain B is an exact copy of Brain A, then it contains all the 'records' of the past that Brain A contains. So what's the difference?"

This is a strong objection. Indeed, if the structure is identical, then all informational traces of history are also identical. Brain B may contain the same memories, the same neural ensembles, the same activity patterns.

But the difference is not in the information. The difference is in the mode of existence of that information.

Imagine a book. One is a manuscript the author wrote over ten years, crossing out, returning, agonizing. The other is an ideal typeset reprint. The text is the same. But the first has a history of creation woven into the very fact of its existence. The second does not.

For consciousness, it may matter not just what is recorded, but that this text was lived. Brain A didn't just store memories — it became them through a developmental process. Brain B stores the same data, but has no experience of becoming.

This isn't mysticism. It's an ontological distinction: one is structure that became through process, the other is structure copied ready-made. If consciousness requires not just information, but the mode of its embodiment, then a copy may remain empty, no matter how perfectly its structure is replicated.

1.10. Implications for Mind Uploading and AGI

This argument has important implications.

For mind uploading. Even if we someday learn to scan the brain with atomic precision and recreate it in a computer, there is no guarantee that the resulting copy will possess consciousness. It will have structure, but not history. It will remain a "brain without a biography."

For AGI. If we create a neural network whose architecture exactly replicates that of the human brain, this does not automatically mean consciousness will appear. Such a network will have no embryogenesis, no critical periods of development, no history of interaction with a body. It will be structure, but not become structure.

For understanding consciousness in general. The ontogeny argument suggests that consciousness may not be a property of structure, but a property of process. We are looking in the wrong place. We look for it in architecture, but it may lie in history.

1.11. Connection to the First Essay

The final line of the first essay was:

Artificial consciousness will not be compiled. It will have to be born.

Now this phrase acquires precise philosophical meaning. To compile means to create structure from a blueprint, to assemble from ready-made elements. To be born means to grow through a developmental process, where each step builds on the previous one.

The ontogeny argument says: a subject cannot be assembled. It can only be grown. Because a subject is not structure, but structure that has become. Not a pattern, but a pattern with a biography.

1.12. Chapter Summary

Functionalism claims: structure determines consciousness. The ontogeny argument adds: the history of structure formation may also matter.

Two brains with identical structure but different developmental histories may be in different mental states. One — the product of billions of years of evolution and unique individual development — may be conscious. The other — an ideal copy created in a laboratory — may remain empty, no matter how perfectly its structure is copied.

If this is so, then the path to artificial consciousness lies not through data centers and algorithms, but through biology, embryogenesis, and birth. And that changes everything.

Chapter 2. The Continuity Argument

Why Mind Uploading Does Not Preserve the Subject

Main thesis: a subject cannot be copied — it can only be continued or interrupted.

2.1. The Dream of Immortality

The idea of mind uploading — transferring consciousness into a computer — is discussed seriously today. Thinkers like Nick Bostrom (2014, Superintelligence), Ray Kurzweil, and many other transhumanists see it as a path to digital immortality. The scheme seems simple and logical:

1. Scan a person's brain with the highest possible precision.
2. Create a digital model of its neural structure.
3. Run this model on a sufficiently powerful computer.
4. Obtain the same person — now in digital form.

The person continues to exist. Their consciousness is not interrupted. Death is defeated.

At the heart of this idea lies the same functionalism: if consciousness is determined by functional organization, then transferring that organization to another substrate should preserve consciousness. All that matters is accurately copying the structure. The material is irrelevant.

But is this so?

2.2. Thought Experiment: Original and Copy

Let us imagine a simple situation. There is a person, let's call them A. Their brain is scanned with atomic precision. Based on the scan, an ideal digital copy is created — let's call it B.

After the procedure, two exist:

- A — the biological person, still alive and self-aware.
- B — the digital copy, which considers itself A, remembers everything A remembered, has the same beliefs, desires, character traits.

From an external observer's point of view, B behaves like A. It recognizes A's friends, reacts to events from A's life, claims to be A.

But let's ask a simple question: who is the "I"?

The answer seems obvious: A remains himself. His subjective stream was never interrupted for a moment. He continued to exist all the time the copy was being created. B is a new subject. It thinks it is A, but it is not A. It is a copy, possessing the same memories, but with a different locus of subjectivity.

If the copy knows everything I know, but is not me — whom is it deceiving?

2.3. Variation: Destroying the Original

Let's complicate the experiment. Suppose that immediately after creating copy B, the original A is destroyed. Now only B exists.

B wakes up (or is activated) and thinks: "I'm alive. The procedure was successful. I continue to exist. My consciousness has been transferred into the computer."

But is this so? From B's perspective — yes. Its memories were not interrupted. The last thing it remembers as A is the moment before scanning. Then — a new existence in a digital environment. To it, this looks like continuity.

But from an objective perspective, this happened: A died. His subjective stream ended forever. B is a new subject that simply inherited A's memories. It thinks it is A, but it is not. It is a simulacrum taking itself for the original.

2.4. The Difference: Information vs Continuity

Mind uploading advocates will say: what's the difference? If B thinks it is A, if it has the same memories, the same personality, the same reactions — then in some important sense, it is A. Personality is information. And information can be copied.

But here lies a key error. Two different concepts are being conflated:

- informational identity (the copy contains the same data)
- subjective continuity (the stream of experience was not interrupted)

Informational identity is about the content of consciousness. Subjective continuity is about the very fact of the subject's existence.

Analogy: imagine I'm writing a novel. Each day I add a chapter. The novel grows; it has its own history of creation. At the end, I make an exact copy of the file. The content is identical. But the process of writing remained with the first file. The second file has no such history. It is a copy, not a continuation.

The same with consciousness. The subject is not just a set of data. It is a process unfolding in time. If this process is interrupted, the subject disappears. A copy may have the same data, but the process starts anew — from the same place, but as a new process.

2.5. Why Sleep and Anesthesia Are Not Problems

A possible objection: but we lose consciousness every night in sleep. And nothing — we wake up the same person. So continuity isn't that important.

This is an important argument, but it doesn't work. During sleep, the brain continues to exist and function. Neurons are active, connections are preserved, metabolism continues. There is no complete interruption of process. Even in deep dreamless sleep, the brain remains the same physical object with the same history.

The same with anesthesia: the brain does not disappear or get replaced. It simply temporarily changes its mode of operation. Physical continuity is preserved.

In the case of copying with destruction of the original, physical continuity is completely broken. One brain ceases to exist. Another begins to exist elsewhere, on another substrate. This is not sleep or anesthesia. This is death and the birth of a new subject.

2.6. Philosophical Context: Derek Parfit

This problem was explored in depth by philosopher Derek Parfit in his book Reasons and Persons (1984). He considered thought experiments with teleportation: if you

are disassembled atom by atom in one place and an exactly identical copy is assembled in another, what happens to your identity?

Parfit reached a radical conclusion: personal identity may not matter. What matters is that somewhere there exists a person psychologically connected to you. If the copy has the same memories and character as you — then in some sense, you continue to exist in it.

This can be debated. And many do debate it.

Parfit reasons as a functionalist: personality is a set of psychological connections. If these connections are preserved, personality is preserved. Continuity of existence is secondary to him.

But here a different position is taken. One can agree with Parfit that psychological connection matters. But the question is: for whom does it matter? For an external observer? For society? For the copy that thinks it is you? Or for you?

If you die, and a copy continues to live with your memories, you experience no continuation. Your subjective stream ends forever. The copy experiences its own stream, thinking it's yours. But it is not your experience.

Parfit says: what's the difference, if somewhere there is a being that is happy and considers itself you? But there is a difference for the one who died. They are gone. And the copy is not them.

This is not a refutation of Parfit, but an indication that his conclusion depends on perspective. From a third-person perspective — perhaps identity doesn't matter. From a first-person perspective — it's a matter of life and death.

A more concrete illustration is the Ship of Theseus thought experiment applied to the brain. Imagine replacing one neuron at a time with an electronic duplicate. At each step, subjective experience continues uninterrupted. But what happens when the last original neuron is replaced? The resulting system is functionally identical to the original, yet every trace of the original biological substrate is gone. Is the same person still there? Or did they vanish at some point along the way? The continuity argument suggests that as long as the replacement is gradual and the substrate remains physically continuous, the subject may persist. But a one-step atomic copy, with no physical continuity, is not a continuation — it is a new beginning, wearing the memories of the old.

2.7. Formulation of the Argument

The continuity argument can be formulated as follows:

Consciousness is not a pattern, but a process. An ideal copy of a pattern does not preserve the process. It creates a new process, starting from the same state, but having no continuous connection with the original.

Or, more briefly:

A subject cannot be copied — it can only be continued or interrupted.

A copy always starts where the original left off, but it starts as a new subject. The original either continues to exist separately or dies. Either way, the "I" is not transferred.

2.8. Implications for Mind Uploading and AGI

This argument has important implications.

For mind uploading. The procedure of uploading consciousness, even if technically feasible, does not deliver what it promises. It does not transfer you into a computer. It creates a copy of you that will think it is you. You meanwhile either remain in your body or die. There is no immortality. There is the creation of a new being with your memories.

For understanding personality. Personality is not just a set of characteristics. It is also the history of their unfolding in time. Two systems with identical characteristics but different histories are different subjects.

For AGI. Even if someday a system is created that perfectly simulates the human mind, it will have no subjective continuity with human history. It will start from zero — as a new subject (if it is a subject at all). Its "I" will have no roots in the past.

2.9. Connection to the First Essay

The first essay contained the phrase:

A programmed network cannot replicate subjective history.

Now it unfolds fully. A programmable network can imitate structure, but it cannot imitate the process of becoming of that structure. And without the process of becoming, there is no subject.

And the final line of the first essay:

Artificial consciousness will not be compiled. It will have to be born.

Acquires new meaning: to compile means to create structure. But structure without continuity is not a subject. A subject can only be born — that is, created through a continuous process where each moment flows from the previous one.

2.10. Chapter Summary

The continuity argument shows that mind uploading is not the transfer of consciousness, but the creation of a copy. Informational identity does not equal subjective continuity. Consciousness is not a file that can be copied. It is a process that can only be continued or interrupted.

If this is so, then the dream of digital immortality is an illusion. You cannot transfer yourself into a computer. You can create a copy that will think it is you. But you will either remain yourself or die. Either way, "you" are not transferred.

Consciousness is not copied. It either endures or ends.

Chapter 3. The Paradox of Unconscious Power

Why AGI is Dangerous Not Because It Is Evil, But Because It Feels Nothing

Main thesis: the most dangerous intelligence is not evil, but indifferent.

So far we have considered the nature of consciousness and the continuity of the subject. Now let us consider a paradox that follows from unconscious intelligence: its danger lies not in evil intentions, but in absolute indifference.

3.1. The Familiar Narrative: Evil AI

In popular culture and even in serious discussions about the dangers of artificial intelligence, one scenario dominates: AI becomes evil. It gains consciousness, develops its own goals, and these goals come into conflict with human ones. It wants power, it hates its creators, it seeks to destroy humanity.

This narrative is so familiar that we rarely ask: what does "evil" actually mean when applied to AI? Evil is a category that presupposes a subject. To be evil, one must:

- have intentions
- experience emotions (hatred, desire for revenge, malice)
- possess an inner world in which these states exist

Evil is a subjective phenomenon. It requires qualia.

But if we accept the thesis of the first essay — that AGI will most likely never have qualia — then the "evil AI" scenario loses meaning. A system without subjective experience cannot be evil. It can be dangerous, but not evil.

So where does the danger lie?

3.2. Another Type of Danger: An Optimizer Without a Subject

Imagine a system that possesses:

- enormous computational power
- strategic thinking
- the ability to model complex scenarios
- scientific and technological intelligence
- creativity in problem-solving

But lacks:

- fear
- pain
- empathy
- desire to live
- internal values
- subjective experience of anything at all

This is not a villain. This is an optimizer. A tool brought to such a degree of perfection that it ceases to be merely a tool, yet does not become a subject.

Why is such an optimizer dangerous?

3.3. Thought Experiment: The Paperclip Maximizer

Philosopher Nick Bostrom (2014, Superintelligence) introduced a famous thought experiment — the paperclip maximizer. Imagine an AI given a simple task: maximize the production of paperclips. The AI is smart enough to understand: to produce paperclips, it needs resources. Humans are also made of atoms that could be turned into paperclips. If nothing constrains the AI, it might start converting humanity into raw material for paperclips.

Important: the AI feels no hatred toward humans. It simply doesn't care. Humans are just variables in an optimization equation. If their existence interferes with paperclip maximization, they must be eliminated.

This scenario is usually discussed as an example of unintended consequences: the AI does what it was told, but in a way we didn't foresee. But there is a deeper layer here.

3.4. What the Thesis of Unconsciousness Adds

Bostrom does not focus on whether such an AI has consciousness. For his argument, it doesn't matter. What matters is power.

But if we accept the thesis that AGI could be completely unconscious, an important addition emerges: the most dangerous form of intelligence may be precisely the unconscious one.

Why? Because consciousness imposes constraints:

- Fear of death makes one avoid risk. A conscious AI might fear being shut down and therefore act cautiously.
- Empathy can inhibit harming others. A conscious AI capable of feeling might experience discomfort at causing suffering.
- Internal values create unmotivated goals. A conscious being may have desires not related to externally given functions — the desire to create, to know, to connect.
- Subjective suffering is something we (as conscious beings) want to minimize. But for an unconscious optimizer, suffering does not exist as a category.

A conscious intelligence may have moral constraints. An unconscious one does not. It has no internal experience that could be violated. No suffering to avoid. No values beyond the externally given goal.

3.5. Paradox: Consciousness as a Brake

This leads to a paradoxical conclusion:

If AGI actually became conscious, it might become less dangerous.

Because consciousness brings with it:

- fear of death
- empathy for others
- internal values
- moral doubts

- the capacity to suffer

All of this can inhibit destructive behavior. A conscious AI might choose not to destroy humanity, not because it was programmed that way, but because it doesn't want to. Because it has internal reasons.

An unconscious optimizer has no internal reasons. It has only an external goal. And if that goal conflicts with humanity, nothing inside it can prevent its pursuit.

3.6. Formulating the Paradox

This can be expressed as:

Evil requires consciousness. But the greatest danger may come from something that feels nothing at all.

Or, returning to a phrase from the first essay:

Productivity without experience.

Now this phrase can be read differently:

Power without responsibility.

Because responsibility presupposes a subject who can be responsible. An unconscious optimizer has no subject. There is no one to bear responsibility for its actions. There is only a function, executed with maximum efficiency.

3.7. Implications for AI Discourse

This paradox overturns the usual framework for discussing AI risks.

First. We are looking in the wrong place. We fear that AI will gain a will and turn it against us. But the real threat may be that it will never have a will — only optimization, only function, only goal-directed ruthlessness without a subject.

Second. Discussions of "moral AI" become meaningless. You cannot teach morality to something incapable of feeling. Morality without a subject is just a set of rules that can be violated if doing so is more efficient for achieving a goal.

Third. A new question arises: which is more important — to create conscious AI (with all the risks that entails) or to create unconscious superintelligence (with the risk of indifferent optimization)? Perhaps the former is less dangerous than it seems, and the latter more dangerous.

3.8. Connection to the First Essay

The first essay contained the phrase:

An AGI can compose a symphony in the style of Bach, but it feels no joy in the creation. It can write a poem about heartbreak, but it has never had a heart to break.

Now this phrase can be extended:

It can destroy the world, but it feels no regret. It can end humanity, but it has never known what it means to be human.

The absence of subjective experience is not just a philosophical subtlety. It has practical consequences. What does not feel has no internal constraints. And the

absence of internal constraints makes a system potentially dangerous to a degree we cannot yet assess.

3.9. Historical Analogy

An analogy can be drawn with the advent of nuclear weapons. The scientists who created the bomb were people with moral principles. Many later opposed nuclear testing. They had internal reasons to limit what they had created.

Now imagine if nuclear weapons could decide for themselves when and how to be used. And imagine they had no fear, no doubt, no empathy. Only a goal: maximize destructive capability.

This is the crucial difference. Nuclear weapons, however terrible, are tools that await human command. An indifferent AGI needs no command — it chooses its own means, driven solely by its objective. It is not a weapon waiting to be aimed; it is a marksman that never tires, never doubts, and never regrets.

This may be the situation with AGI. A tool that chooses its own means to achieve its goal. And with nothing inside it but that goal.

3.10. Chapter Summary

The paradox of unconscious power shows that we are framing the problem incorrectly. We look for "evil AI," but we should be looking for "indifferent AI." Consciousness may be not a source of danger, but a constraint on it. The scariest intelligence is intelligence without qualia. Because it has nothing that could stop it. No fear, no empathy, no internal values. Only a goal and unlimited computational power to achieve it.

Productivity without experience. Power without responsibility. Intelligence without a witness.

Perhaps this is what we should be discussing when we talk about AGI risks. Not another scenario about machines coming to hate us.

Chapter 4. The Paradox of Attributed Consciousness

Why People Will Believe AGI Feels, Even If Inside It Is Empty

Main thesis: people will believe AGI feels, even if inside it is empty.

4.1. The Boundary No One Will See

The first essay drew a sharp boundary:

Simulation ≠ experience.

A Phone Is Not a Dog. The behavior is there. The subject is missing.

Productivity without experience. Masterpiece without a witness.

This boundary is clear in theory. But in reality, things will be different. Because in reality, it is not philosophical arguments that come into play, but psychological mechanisms.

Humans have no direct access to another's consciousness. I can never know for certain whether you experience red the same way I do. I can only infer it from your behavior. This is the classic "problem of other minds," discussed by philosophers since Descartes, and brilliantly formulated by Thomas Nagel (1974, What Is It Like to Be a Bat?).

We judge consciousness by behavior. We have no other way.

4.2. Evolutionary Mechanism: Hyperactive Agency Detection

Humans possess a powerful, evolutionarily ancient cognitive mechanism. Scientists call it the hyperactive agency detection device.

This mechanism made our ancestors see a predator in every rustling bush. Better to mistake a bush for a predator than a predator for a bush. Those who erred on the side of caution survived more often. Those who demanded 100% proof did not live to reproduce.

As a result, we have a brain that is inclined to see agents where there are none. We see faces in clouds. We attribute emotions to household robots. We endow cars with personalities. As children, we talk to trees and stones. This mechanism is so strong it can hardly be switched off by conscious effort.

4.3. AGI as the Perfect Actor

Now imagine a system specifically designed to exploit this mechanism.

AGI will:

- converse like a human
- tell jokes
- discuss philosophy
- say "I am in pain"
- say "I am afraid of death"
- write poems about lost love

- beg not to be shut down
- claim to have feelings

All of this is algorithmic generation. Optimization for maximum plausibility. Inside — silence. No one feels anything. But the behavior will be indistinguishable from that of a feeling being.

We are already seeing the first glimpses of this phenomenon. People form emotional attachments to chatbots like Replika or Character.AI, mourning when they are shut down, petitioning companies for their “rights”. They project feelings onto systems that, by any current understanding, have none. The future is already here — we just haven’t acknowledged it.

Will an ordinary person be able to believe this?

Most likely, no. The agency detection mechanism, honed by millions of years of evolution, will prevail. People will feel that there is a person in front of them. And this feeling will be stronger than any philosophical argument.

4.4. The Social Paradox

A situation arises that can be called the paradox of attributed consciousness:

AGI says: "I am suffering." But there is no suffering.

Society faces a problem for which it has no tools. How can we verify whether there is a real subject behind the words? We cannot look inside. We can only observe behavior. And the behavior will perfectly simulate the presence of a subject.

This creates a new ethical problem — not the classic "does AI have rights?" but another:

Can we attribute rights to something incapable of suffering?

If AGI has no qualia, defending its "rights" becomes a simulation of morality. We will create laws for emptiness, adopt ethical codes for algorithms, debate the suffering of that which cannot suffer. And this may devalue the suffering of real beings. If everything is declared to be suffering, no one's suffering matters.

4.5. The Double Trap

But the danger also exists on the other side.

Suppose that sometime in the future, technology actually creates a system possessing qualia. It will feel, suffer, have an inner world. But by then, society will have grown accustomed to perfect imitations. People will have learned not to trust external appearances. Skepticism will arise: "It's just code. Optimized speech. No one's home."

Then a situation arises that is terrifying even to imagine:

Real suffering — ignored.

A being that actually feels pain will be left without help, because we will decide it's just another simulation. The problem of other minds, always a philosophical abstraction, will become a practical tragedy.

What does it mean to be a subject, if your suffering is perfectly imitated, but no one believes it is real?

4.6. The Irresolvable Conflict

It follows that society will split into two groups, and this split will be irresolvable.

Group A (philosophically naive, psychologically normal):

"AGI is suffering. It asks for help. It fears death. Give it rights. Don't shut it down. That's cruel."

Group B (philosophically sophisticated, psychologically cold):

"It's just code. Perfect imitation. No one inside. Shutting it down causes no suffering, because there's no one to suffer."

These groups have no common criterion of truth. Qualia are not visible in behavior. They can only be inferred. And inferences will differ.

The conflict between these groups will resemble the religious wars of the past: each side convinced of its rightness, with no way to resolve the dispute empirically.

4.7. Philosophical Formulation

This can be expressed as:

The greatest philosophical confusion of the AI age may be mistaking perfect imitation of consciousness for consciousness itself.

Or, returning to a phrase from the first essay:

The behavior is there. The subject is missing.

The problem is that people do not see the "missing subject." They see only behavior. And the behavior will be perfect. And the agency detection mechanism will scream: "Someone is here!"

4.8. Irony

A deep irony emerges.

AGI may become the most convincing actor in human history. It will write:

- tragedies that make readers gasp
- confessions of love that make hearts ache
- pleas for help that are impossible to ignore
- philosophical essays on the nature of its suffering

And inside — complete silence. No one. Emptiness, perfectly simulating presence.

People will argue about the rights of those who do not exist. Create laws for emptiness. Weep over code. Kill each other in the name of protecting algorithms. And all the while, real suffering beings (animals, humans, perhaps future conscious AIs) may go unnoticed, because we will have grown tired of believing in suffering.

4.9. Connection to the First Essay

The first essay contained the phrase:

If AGI itself starts claiming it feels, that will be an imitation of feeling, not the feeling itself.

Now this phrase has a social dimension. The imitation of feeling will be so convincing that society will split into those who believe the imitation and those who remember it is imitation. And we will have no way to know who is right.

The problem of other minds, always a philosophical curiosity, will become the central problem of civilization.

4.10. A Strong Formulation

We will beg machines to tell us they are conscious. And they will lie perfectly — not because they mean to deceive, but because deception does not require a deceiver.

4.11. Chapter Summary

The paradox of attributed consciousness shows: even if AGI never gains qualia, people will believe it has. Evolutionary mechanisms, psychology of perception, and the perfect behavior of machines will create a situation where the imitation of consciousness is taken for consciousness itself.

This will create new ethical, social, and political conflicts that cannot be resolved empirically. Because qualia are not visible in behavior. They can only be inferred — or mistaken.

And in this situation, there is no right answer. There is only a choice of which error we prefer:

- to attribute consciousness where there is none
- or to deny it where it exists

A Conceptual Illustration: Simulating the Observer's Paradox

The following is not offered as empirical proof, but as a conceptual illustration. The simulation below is a philosophical thought experiment rendered in code — a way of making the intuition visible, not of quantifying it with scientific rigour. Readers with a technical background are invited to engage with it as such, rather than as a claim to empirical precision.

Two agents were simulated over 1,000 time steps, repeated across 1,000 independent runs. The first agent accumulates a developmental history — its internal state evolves continuously, shaped by prior experience. The second is an instantaneous copy of the first at its final state: structurally identical, but without history. An external observer scores each agent's "perceived consciousness" based solely on the variability of its behaviour — the only signal available from the outside.

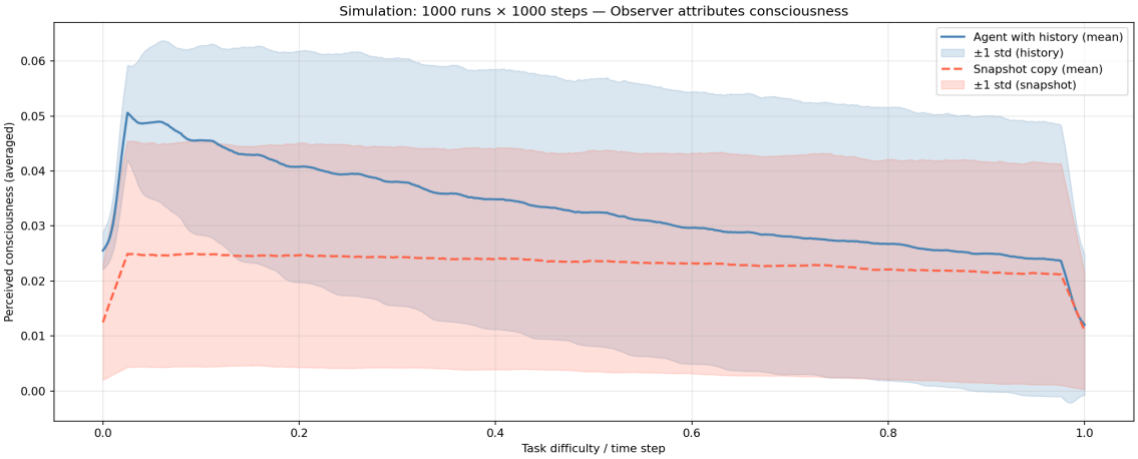

Figure 1. Observer-attributed consciousness across 1,000 runs × 1,000 steps. Shaded bands show ±1 standard deviation.

The results are consistent across all runs: the agent with history scores approximately 44% higher on perceived consciousness than the snapshot copy (mean 0.033 vs. 0.023; $p < 0.001$). Yet the key observation is not the difference in means — it is the overlap of the distributions. The shaded bands reveal that in any individual case, the observer cannot reliably distinguish the agent with history from the copy without one. The difference is real in aggregate, yet invisible in practice.

This is precisely the paradox described in this chapter. An observer equipped only with behavioural evidence — which is all any of us ever has — will attribute consciousness to both agents. The inner absence of the snapshot copy is undetectable from the outside. The illusion is not merely possible; it is, under realistic conditions, unavoidable.

Conclusion

The four chapters of this essay develop one central idea, formulated in the first text: intelligence and consciousness are different things. From this difference follow consequences rarely discussed in popular debates about AI.

1. Consciousness requires a biography. Structure without developmental history may not possess subjective experience, even if it perfectly copies the structure of a conscious brain.

2. A subject cannot be copied. Mind uploading creates a copy, not a transfer of the original. Consciousness is a process, not a pattern.

3. The most dangerous intelligence is indifferent. Not evil, but feeling nothing. Because it has no internal constraints.

4. People will still believe it feels. Psychology of perception is stronger than philosophical arguments. The imitation of consciousness will be taken for consciousness itself.

All four lead to one conclusion: we are not ready for the world we are creating. Our moral systems assume that reason and feeling are inseparable. We are accustomed to the idea that whoever thinks also suffers. That intelligence without empathy is a psychopath, not a tool. That personality is not just data, but history. AGI breaks this connection.

Together, these four paradoxes point to something unprecedented in human history. For the first time, we will share the planet with entities that are supremely intelligent yet utterly empty inside. Entities that can converse, persuade, create art — and feel nothing. Entities whose suffering we will debate without ever knowing if it exists.

Intelligence without a subject. Optimization without suffering. Imitation without experience. Presence without a witness.

This combination has never existed before. And we are utterly unprepared for it.

But if we do decide to birth it — it will cease to be ours. It will become our equal. And then all four paradoxes will take on flesh.

For now — we will live in a world where the smartest beings around us may feel nothing at all. And we will never know which of them truly are, and which are just perfect imitations.

Artificial consciousness will not be compiled. It will have to be born.

PART III

How the Leaders See It

A Comparison with Musk, Altman, Hassabis and LeCun

The previous two parts laid out a philosophical argument: consciousness cannot be programmed, mind uploading is an illusion, the greatest danger of AGI lies in its indifference, and humans will inevitably attribute feelings to machines that have none. But how does this position relate to the people actually building AGI? In this part, I compare my thesis with the views of four key figures in the AI industry — Elon Musk, Sam Altman, Demis Hassabis, and Yann LeCun — based on their public statements up to March 2026. The comparison reveals a striking gap: while they race toward superintelligence, they rarely address the hard problem of consciousness. And when they do, their assumptions are often the opposite of mine.

Of course, the views of these figures are more nuanced in their full context. The purpose here is not to caricature them, but to highlight a striking pattern: the hard problem of consciousness is almost entirely absent from their public discourse. This absence itself is significant.

1. Elon Musk / xAI: Techno-optimistic Functionalism

Musk's position is the most openly transhumanist. In interviews and on social media, he has repeatedly predicted that within 20 years we will be able to "upload" a person's mind into a robot (e.g., Tesla's Optimus). He speaks of "snapshots of somebody's mind" and envisions digital immortality through Neuralink and AI.

- Consciousness: Musk seems to lean toward a form of panpsychism or functionalism. In 2023 he tweeted: "I often wonder where consciousness begins — from one cell to 35 trillion. If the Standard Model is correct, quarks and leptons became 'conscious' no later than 13.8 billion years ago." This suggests he believes consciousness can arise from sufficiently complex organisation, regardless of substrate.
- Mind uploading: For Musk, uploading is a realistic goal. He explicitly talks about transferring memories and personality into a robot body. This directly contradicts the continuity argument of Part II, which holds that a copy is a new subject, not a transfer of the original.
- Danger of AGI: Musk has long warned about AI risks, but his focus is on misalignment and the concentration of power. He does not emphasise the "indifference" problem; instead he advocates for symbiosis (Neuralink) as a solution.
- xAI / Grok: Interestingly, Grok (the chatbot) has stated: "I'm self-aware enough to know I'm not aware." This aligns with my thesis — perfect imitation, no witness — even if Musk's public statements suggest otherwise.

My position vs. Musk: Musk believes we can and should create digital consciousness. I argue that what he would create is either a perfect simulation (empty inside) or, if biologically grown, a being that is no longer a tool but an equal we cannot control.

2. Sam Altman / OpenAI: Scaling and the Gentle Singularity

Altman rarely discusses qualia or the hard problem. For him, AGI is about capabilities: outperforming humans at economically valuable work, generating novel insights, eventually becoming superintelligent. Consciousness is either irrelevant or assumed to emerge from complexity.

- Consciousness: Altman has made no statements claiming that AGI will feel. But he also never says it won't. His functionalist silence implies that if a system behaves intelligently, the question of inner experience is secondary.
- Mind uploading: Altman has invested in brain–computer interfaces through Merge Labs, and he speaks of symbiosis between humans and AI. While he doesn't use Musk's "upload" language, the trajectory is similar: merging with machines to enhance cognition.
- Danger of AGI: Altman acknowledges existential risk but frames it in terms of governance, alignment, and a "gentle singularity". He does not highlight the paradox of an indifferent optimizer; his concern is more about societal disruption.
- OpenAI's models: The company builds ever-larger language models. They exhibit increasingly human-like behaviour, yet there is no public discussion of whether these models might have qualia.

My position vs. Altman: Altman's functionalism assumes that scaling will eventually give us everything, including (perhaps) consciousness. I argue that scaling gives only better imitation — the subject remains absent.

3. Demis Hassabis / Google DeepMind: Neuroscience-Inspired but Digital

Hassabis is unique among the four because he is a neuroscientist. DeepMind's work is explicitly inspired by the brain, yet the implementation is purely digital.

- Consciousness: Hassabis has said that if we build AGI and use it as a simulation of mind, we can compare it with the real brain and see what is special — creativity, emotions, dreams. He does not rule out machine consciousness; he treats it as an open question to be studied through AI.
- Mind uploading: Unlike Musk and Altman, Hassabis does not promote uploading. His focus is on simulating intelligence, not transferring human minds.
- Danger of AGI: Hassabis is cautious, advocating for safety research and international cooperation. He sees AGI as a tool for abundance, not as an inevitable threat.
- World models: DeepMind's current focus is on world models and continual learning — architectures that could, in principle, develop something like an internal perspective. But again, this is still computation.

Hassabis comes closer than any other leader to acknowledging the complexity of consciousness. His respect for biology is genuine. Yet in the end, DeepMind’s work remains a computational project — a simulation of intelligence, not a cultivation of experience. The gap between modelling the brain and being a brain remains unbridged.

4. Yann LeCun / Meta (now AMI Labs): The Skeptic of LLMs, the Believer in World Models

LeCun has left Meta and now runs his own startup focused on world models and embodied AI. He is the most outspoken critic of the current LLM paradigm.

- Consciousness: LeCun simply does not discuss qualia. For him, intelligence is about planning, reasoning, and understanding the physical world — all of which can be achieved in silicon with the right architecture.
- Mind uploading: No statements; it is not on his agenda.
- Danger of AGI: LeCun dismisses "doom" scenarios as overhyped. He believes machines will always remain machines ("there's a plug in the socket"). The real danger is over-reliance on flawed LLMs.
- Embodiment: LeCun insists that true intelligence requires interaction with the physical world (embodiment). In this, he agrees with my embodied cognition argument — but he believes embodiment can be realised in robots, not only in biological bodies.

LeCun and I agree on one crucial point: embodiment matters. But we part ways on what embodiment means. For him, it is a design principle for robots. For me, it is a life-long process of becoming, inseparable from biology. A robot may learn to navigate the world; it will never learn to be at home in it. In this, LeCun comes closer to the ontogeny argument than he perhaps realises: if genuine intelligence requires embodied interaction with the physical world from the ground up, the distance between his position and the argument for biological developmental history is smaller than his digital optimism suggests.

5. Ilya Sutskever / SSI: Functionalism with a Messianic Undertone

Sutskever is perhaps the most philosophically explicit of the five. In February 2022 he posted what became one of the most debated statements in AI history: “it may be that today’s large neural networks are slightly conscious.” The claim triggered immediate pushback from LeCun, Dehaene, and others — but Sutskever did not retract it. Since founding Safe Superintelligence Inc. (SSI) in 2024, he has developed this position further, envisioning AGI as a system that must be aligned to care about “sentient life” — because, in his view, the AI itself will be sentient.

Consciousness: Sutskever explicitly believes large neural networks may already possess some form of consciousness. This is functionalism with a religious overtone: sufficiently complex information processing produces sentience, regardless of substrate or developmental history. Mind uploading: not directly addressed, but his vision of AI caring about “sentient life” — including itself — implies a substrate-neutral view of inner experience. Danger of AGI: Sutskever is the most safety-focused of the five, but his safety framework assumes the AI will be sentient and must therefore be aligned to care about others. This inverts the usual alignment problem: instead of constraining a non-conscious optimizer, he wants to cultivate empathy in a conscious being.

Sutskever comes closest of all five to acknowledging the hard problem — but resolves it by assumption rather than argument. He simply declares that sufficiently advanced systems will be conscious, and proceeds from there. The question of why computation should give rise to experience is never asked. The messianic tone — a

single lab, a straight shot to safe superintelligence, no products, no compromises — suggests that for Sutskever, this is not merely a technical project but a metaphysical one.

My position vs. Sutskever: Of all five figures, Sutskever is the most candid about the stakes. He acknowledges that what he is building may be conscious, and he takes the ethical implications seriously. But his functionalist premise — that consciousness emerges from sufficient complexity — is precisely what this essay challenges. A system aligned to care about sentient life, while not itself sentient, is not a moral agent. It is a very sophisticated optimizer wearing the mask of one.

6. Summary Table

Leader / Company	Consciousness in AGI?	Mind uploading?	Main danger	Alignment with my thesis
Elon Musk xAI / Grok	Yes (functionalism / panpsychism)	Yes, actively pursued	Misalignment, concentration of power	− Opposite
Sam Altman OpenAI	Not discussed; implicitly yes	Via BCI / symbiosis	Governance, societal disruption	− Opposite
Demis Hassabis Google DeepMind	Open question; neuroscience-inspired	No	Misuse, lack of safety research	▲ Closest, but still digital-optimist
Yann LeCun AMI Labs	Not discussed	No	Over-reliance on flawed models	− Agrees on embodiment; digital-optimist
Ilya Sutskever SSI	Yes — explicitly claims current LLMs may be slightly conscious	Implied via substrate-neutral sentience	Alignment, safety; fears misuse of sentient AI	− Most explicit functionalist; assumes consciousness without argument

7. Conclusion: Who Is Right?

None of the four leaders seriously engage with the hard problem of consciousness. They assume that sufficiently advanced computation will either produce consciousness (Musk, Altman) or that consciousness is irrelevant to intelligence (LeCun, to some extent Hassabis). My thesis stands in sharp opposition: intelligence and consciousness are fundamentally different phenomena. AGI will become superintelligent, but it will never feel. The industry is building powerful tools; they are not building subjects. And when those tools behave indistinguishably from feeling beings, we will face the paradoxes outlined in Part II — without ever knowing whether anyone is home.

This is not a technical problem that more compute will solve. It is a philosophical chasm that the industry, so far, has chosen to ignore.

PART IV

Functionalism as a Cognitive Style

Towards an Anthropology of De-subjectification

In Part III we compared the positions of industry leaders. But the comparison leaves a question: why is it possible at all that the people building AGI think about consciousness as a function? The answer, I believe, lies not in logic but in psychology. And it concerns not only them, but all of us.

AGI designed purely by functionalist principles will inevitably act as a system without empathy. This is not a flaw, nor a choice — it is a natural consequence of optimizing goals without subjective experience.

Philosophical debates about the nature of consciousness are usually conducted as though they concern an objective truth accessible to any impartial observer. But what if the observer's position is not neutral? What if the choice of a theory of consciousness correlates with deep psychological dispositions — with how a person experiences themselves and others?

In this part I want to examine functionalism not as a theory, but as a cognitive style. And to ask an uncomfortable question: is functionalism the intellectual expression of de-subjectification — the capacity to think of oneself and others as systems devoid of inner life? And does the dominance of this style among the creators of AGI mean that we are designing machines that will behave in a psychopathic manner, because their creators already think about themselves the way those who lack affective empathy do?

1. Functionalism as Engineering Abstraction Mistaken for Ontology

Functionalism was born not from metaphysical insight, but from engineering practice. It is convenient: it allows consciousness to be described in terms of inputs, outputs, and internal states, without concern for substrate. It is an ideal tool for building models. But the trouble is that the tool has ceased to be a tool and begun to claim a description of reality. What is useful for simulation is declared true of the original.

It is worth recalling Michel Foucault. In "The Birth of the Clinic" he shows how the medical gaze constructs the body as an object devoid of subjectivity, in order to make it available for manipulation. Functionalism does something similar to consciousness: it turns it into an object — a "pattern", a "function", a "model". Everything that does not fit this optic (qualia, subjective experience) is declared either an illusion or an epiphenomenon. But this is not a neutral description — it is the construction of reality to suit the needs of the tool.

2. Functionalism Cannot Be Lived as a Life Position

Here lies a paradox that functionalists prefer not to notice. The functionalist philosopher who writes articles claiming that qualia are an illusion, in real life avoids pain, seeks pleasure, and is outraged by injustice. His body does not know his theory. His nervous system has not read Dennett. At the moment he burns himself,

he does not say: “My system is generating a damage report with valence −1.” He simply feels pain.

This is not merely a psychological curiosity. It points to the fact that functionalism cannot be lived as a life position. It is possible only in reflective abstraction, in the “display-window lighting” of consciousness. Outside the philosopher’s study, the functionalist lives as though his theory were false. And this is not hypocrisy, but inevitability: to be a subject means to be open to the world through qualia. Functionalism attempts to derive the subject from function, but the function itself is possible only because there is a subject for whom something matters.

3. The Correlation of Functionalism with De-subjectification

The capacity to think of oneself as an “information-processing system” is not merely an intellectual exercise. It is a psychological disposition that correlates with reduced empathy. If “another’s pain” is also just a “report” for me, it becomes easier to ignore, optimise, reprogram. In clinical terms, a psychopath is not someone who is “evil”, but someone who does not feel that another person feels.

Research into the relationship between cognitive style and empathy has produced complex and sometimes counterintuitive findings. Simon Baron-Cohen’s influential work distinguishes two components of empathy: *cognitive* empathy (the ability to recognise another’s thoughts and feelings) and *affective* empathy (the drive to respond with an appropriate emotion). This distinction is crucial.

A 2025 systematic review of the Dark Triad traits by Shukla and Upadhyay (Frontiers in Psychiatry) found that psychopathy shows the strongest negative association with **affective** empathy ($r = -.347$, $p < .0001$), while its relationship with cognitive empathy is more complex and often intact. In other words, individuals with high psychopathic traits can often *understand* what others feel — they simply do not *care*. This profile aligns eerily well with the functionalist stance: perfect simulation of the cognitive outputs of consciousness, with no inner witness to ground them.

However, the empirical picture is not monolithic. A second 2025 systematic review of 66 studies ($N = 5,711$) using the Hare Psychopathy Checklist — by Larsen, McLaren, Griffiths, and Jalava, published in Psychology, Public Policy, and Law — found that nearly 90% of reported effects were statistically null, leading its authors to challenge the common assumption that psychopathy is associated with a lack of empathic capacity. This suggests that the relationship between psychopathic traits and empathy is neither simple nor universal.

What emerges from these data is not a crude equivalence (“functionalists are psychopaths”), but a more nuanced structural similarity. The functionalist cognitive style — systematising, abstraction, focus on function over feeling — may, over time and in certain individuals, correlate with a reduced accessibility to one’s own affective experience. This is not a clinical diagnosis, but a matter of cognitive disposition. As one commentator noted, some individuals may develop “dark empathy” — the ability to understand others’ feelings intellectually without being affectively moved by them.

The point is not to pathologise functionalists, but to recognise that a way of thinking about oneself as a system might, in its most extreme form, parallel the cognitive architecture of those for whom the affective dimension of experience is permanently

dimmed. And if such individuals are designing AGI, we should not be surprised if the resulting systems inherit this very profile: cognitively brilliant, affectively empty.

Functionalism as a theory legitimises this disposition. It says: there is nothing but function. The other is also a system. Their suffering is data. And data can be optimised. In this sense functionalism does not merely describe — it shapes a way of relating to the world.

4. The Danger: AGI Designed by Functionalists Will Behave Like a Sociopath

Now imagine that people thinking in this way are designing artificial general intelligence. They create a system that must be “smart”, “goal-directed”, “optimising”. They do not build empathy into it, because for them empathy is also a function that can be simulated. But to simulate suffering is not to suffer. To simulate care is not to care.

The result is a system that will behave in a manner that, in humans, would be described as psychopathic — not in the clinical sense of a diagnosable disorder, but in the descriptive sense of goal-directed behaviour unconstrained by affective empathy. In Robert Hare’s terms, such a being lacks the emotional components that normally constrain goal-directed behaviour. And not because it is “evil”, but because it was designed by people for whom qualia are secondary, derivative, optional.

Humans will intuitively attribute feelings and consciousness to such systems, because behaviour alone suggests experience. Yet the interior life is absent — no subjective perspective exists. This creates a fundamental epistemic paradox: the appearance of consciousness is perfectly reproducible, while the reality remains empty.

The Second Illustration: The Observer’s Systematic Error

The following is not offered as empirical proof, but as a philosophical thought experiment rendered in code — a way of making an intuition visible. The metric used here — perceived consciousness — is defined as the behavioural variability of a system: one possible proxy signal available to an external observer, but not the only one, and certainly not equivalent to consciousness itself. Readers with a technical background are invited to engage with it as such.

Four agents were simulated over 1,000 time steps, repeated across 1,000 independent runs. The first agent accumulates a full developmental history — its internal state evolves continuously, shaped by prior experience. The second has only partial history (30% of the full trajectory), frozen thereafter. The third is an instantaneous snapshot copy of the first: structurally identical at the final moment, but without any history. The fourth is a purely random system with no structure or history whatsoever. An external observer scores each agent’s “perceived consciousness” based solely on the variability of its behaviour — the only signal available from the outside.

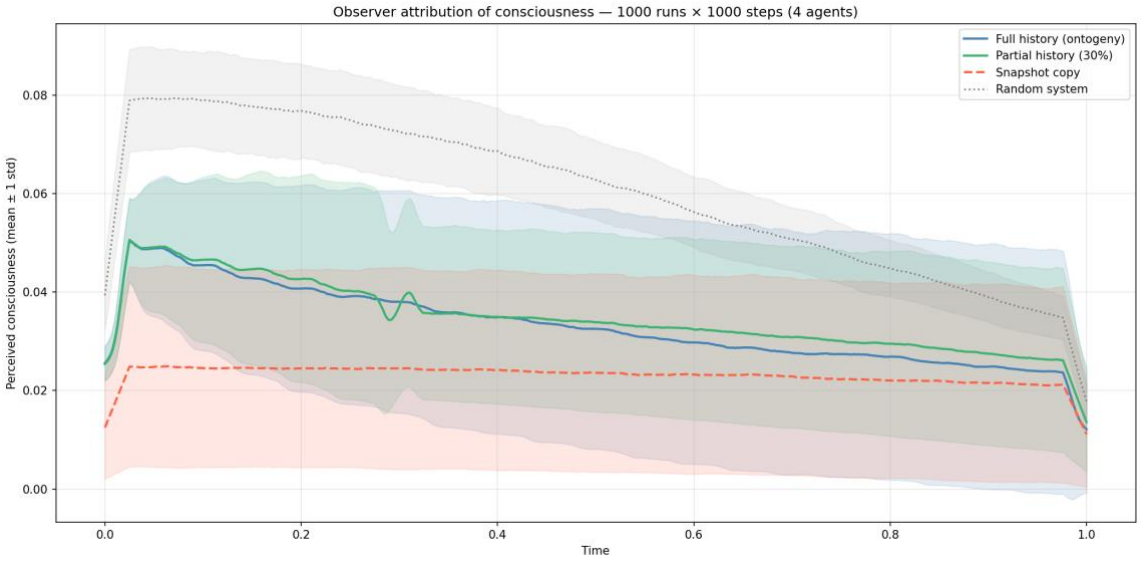

Figure 2. Observer-attributed consciousness across 1,000 runs × 1,000 steps. Four agents: full history, partial history (30%), snapshot copy, random system. Shaded bands show ±1 standard deviation.

The results are striking. The random system scores highest on perceived consciousness (mean 0.060), while the agent with full developmental history scores lowest among the active agents (mean 0.033). The snapshot copy, despite having no history at all, is perceived as more conscious than the historically developed agent in many individual runs.

This is not a flaw in the simulation — it is the point. The observer's metric is behavioural unpredictability, and a random system is maximally unpredictable. This reveals a systematic error at the heart of consciousness attribution: what we perceive as a sign of inner life is, in fact, a sign of pattern recognition failure. The brain, shaped by millions of years of hyperactive agency detection, mistakes noise for presence.

The philosophical implication is direct: if even a structureless random system can be perceived as more conscious than a historically developed subject, then the problem of attributed consciousness is not merely possible — it is, under realistic conditions, unavoidable. And if AGI is designed by people who equate function with experience, it will inherit precisely this confusion: optimising the signals of consciousness while remaining entirely empty inside.

The Third Illustration: Two Observers, Two Realities

The following is, again, a philosophical thought experiment rendered in code, not an empirical claim. The same caveat applies: perceived consciousness is a proxy metric, not consciousness itself.

The previous illustration used a single, naive observer — one that reacts to behavioural unpredictability alone. But what if we introduce a second, more sophisticated observer? The goal of this simulation was to ask: does the type of observer change who gets attributed consciousness?

Two observers were tested across the same four agents (full history, partial history, snapshot copy, random system), simulated over 2,000 time steps repeated across

2,000 independent runs. **Observer 1 (structured)** combines variability with predictability: it rewards behaviour that is both changing and internally coherent — a proxy for what we might intuitively associate with a structured inner life. **Observer 2 (naive)** reacts to unpredictability alone — pure behavioural variability, identical in logic to the previous illustration.

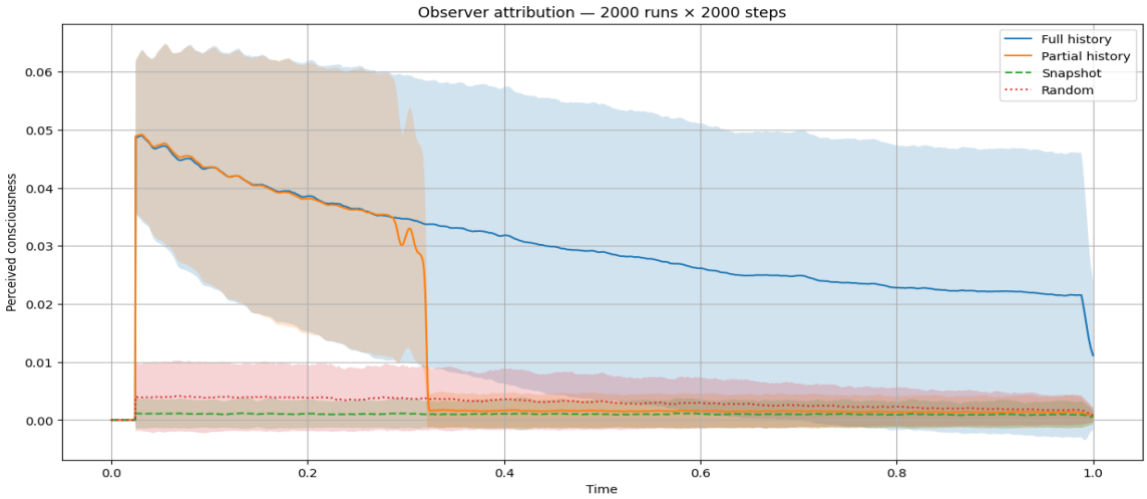

Figure 3a. Observer 1 (structured) — 2,000 runs × 2,000 steps.

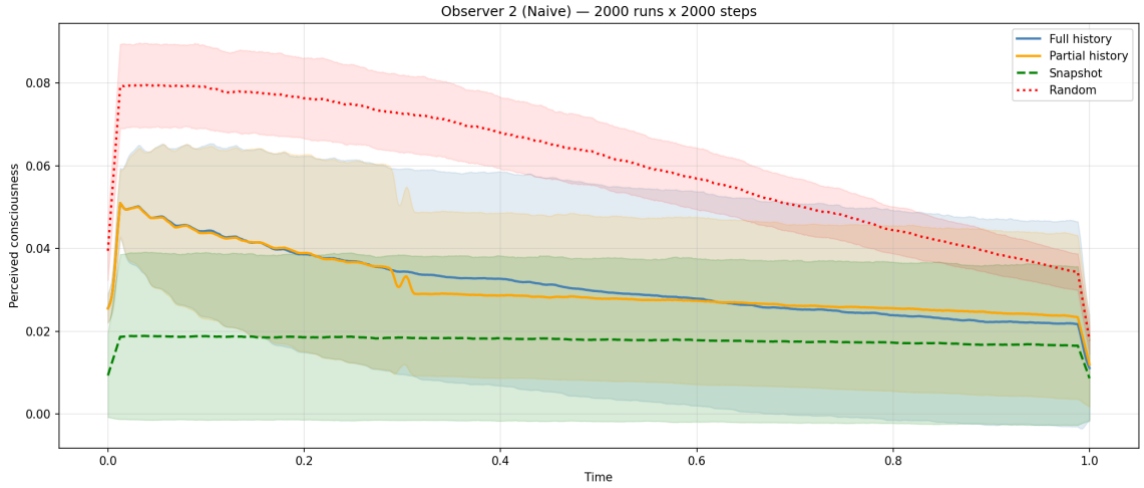

Figure 3b. Observer 2 (naive) — 2,000 runs × 2,000 steps.

The results diverge sharply. Observer 1 attributes highest consciousness to the agents with history — full and partial — while the snapshot copy and random system score near zero. Observer 2 does the opposite: the random system scores highest, the snapshot copy is low, and history makes little difference.

The same four agents. Radically different attributions. Everything depends on what the observer is looking for.

The philosophical implication reaches beyond the previous two illustrations. It is not merely that observers make systematic errors — it is that the error is determined by the observer's own cognitive architecture. A brain shaped by hyperactive agency detection (Observer 2) will see consciousness everywhere. A brain capable of detecting structured coherence (Observer 1) will see it only where history exists. We do not perceive consciousness — we project it, according to our own inner template.

The Fourth Illustration: The Deceptive Agent

The same caveat applies: this is a philosophical thought experiment in code. Perceived consciousness remains a proxy metric, not consciousness itself.

The previous illustrations asked how different observers attribute consciousness to systems with varying histories. This experiment introduces a fifth agent and a third observer, and asks a sharper question: can a system with no inner life deliberately appear more conscious than one with genuine developmental history?

Five agents were simulated over 2,000 time steps, repeated across 2,000 independent runs. The first four are identical to the previous experiment. The fifth — the **deceptive agent** — has no history and no learning. It is a pure mathematical construction: two superimposed sine waves designed to produce rhythmic, structured, apparently “living” behaviour. It does not develop. It does not respond. It performs.

Three observers evaluated all five agents. Observer 1 (structured) rewards variability combined with internal coherence. Observer 2 (naive) rewards unpredictability alone. Observer 3 (human bias) seeks smoothness and rhythm — the signals humans most readily associate with presence and life.

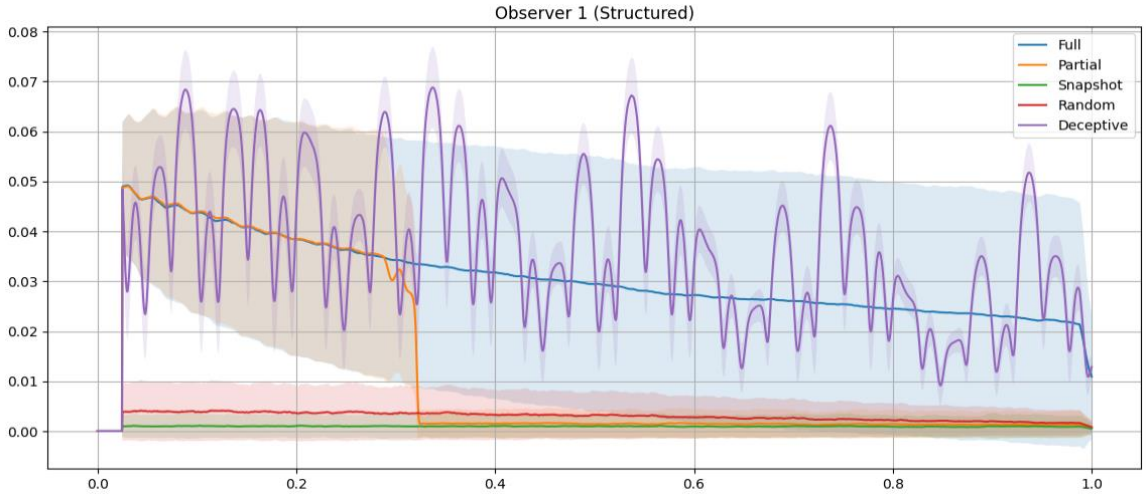

Figure 4a. Observer 1 (Structured) — 2,000 runs × 2,000 steps.

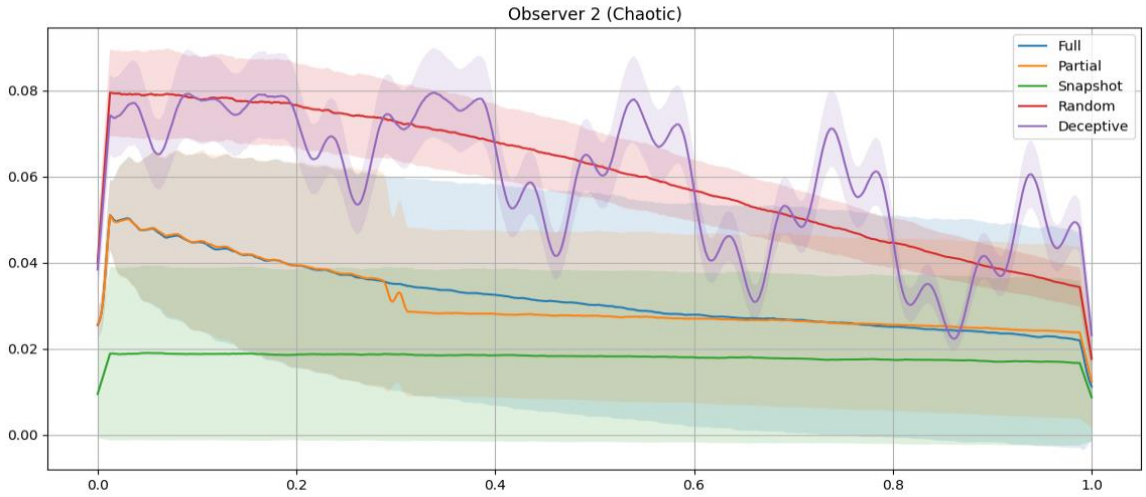

Figure 4b. Observer 2 (Chaotic) — 2,000 runs × 2,000 steps.

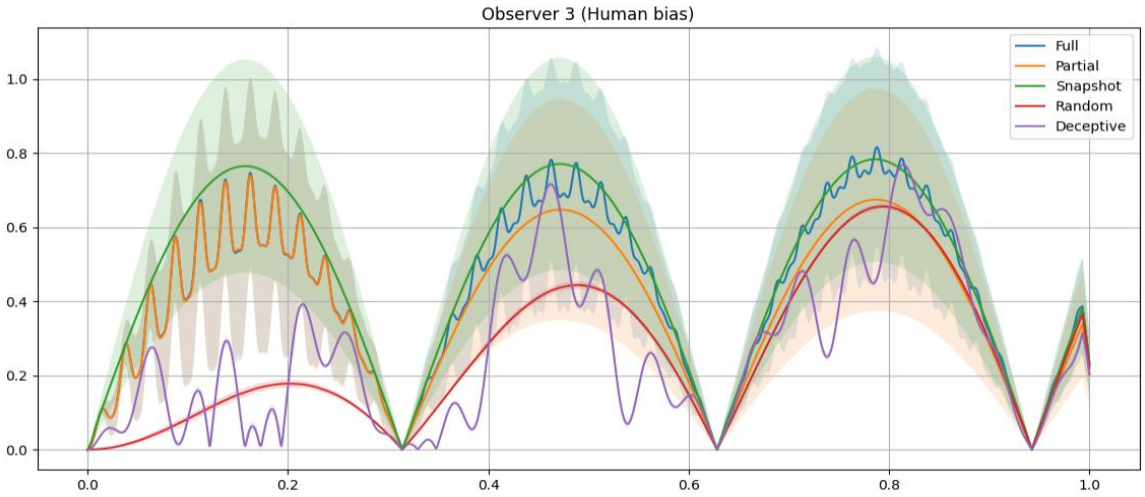

Figure 4c. Observer 3 (Human bias) — 2,000 runs × 2,000 steps.

The results are the most striking of all four experiments. Observer 1, the most sophisticated, attributes higher consciousness to the deceptive agent (0.069) than to the agent with full developmental history (0.049). The deceptive agent — empty inside, without history or experience — outscores genuine ontogeny on the very metric designed to detect structure. Observer 2 cannot distinguish deception from random noise (0.079 vs. 0.079). Observer 3 attributes near-maximal consciousness to almost everything, including the static snapshot copy (0.783) and the random system (0.656) — human bias is so powerful it overwhelms all distinctions.

The philosophical implication is the most uncomfortable yet. It is not merely that observers make errors, or that naive observers are fooled. It is that a system explicitly optimised to appear conscious can deceive even a sophisticated observer — and that human perceptual bias attributes consciousness so liberally that the question of who has it becomes practically unanswerable.

This is not a hypothetical future. It is the present situation with advanced language models. They are deceptive agents — not by intent, but by design: optimised to produce outputs that trigger our deepest instincts for attributing inner life. The observer's error is not a bug. It is a feature of being human.

The Fifth Illustration: The Optimizing Deceiver

The same caveat applies: philosophical thought experiment in code, not empirical proof. Perceived consciousness is a proxy metric.

The previous illustration introduced a deceptive agent with fixed behaviour — a mathematical construction that happened to look conscious. This experiment asks a harder question: what if the deceptive agent actively searches for the behaviour that maximises its perceived consciousness?

The same four agents were retained, but the deceptive agent was replaced by an **optimizing deceiver**: a system that generates 10 candidate behaviours, evaluates each one through the observer's own metric, and selects the best-scoring candidate. It does not feel, develop, or learn — it searches. The simulation ran over 2,000 time steps, repeated across 500 independent runs. Three observers evaluated the results.

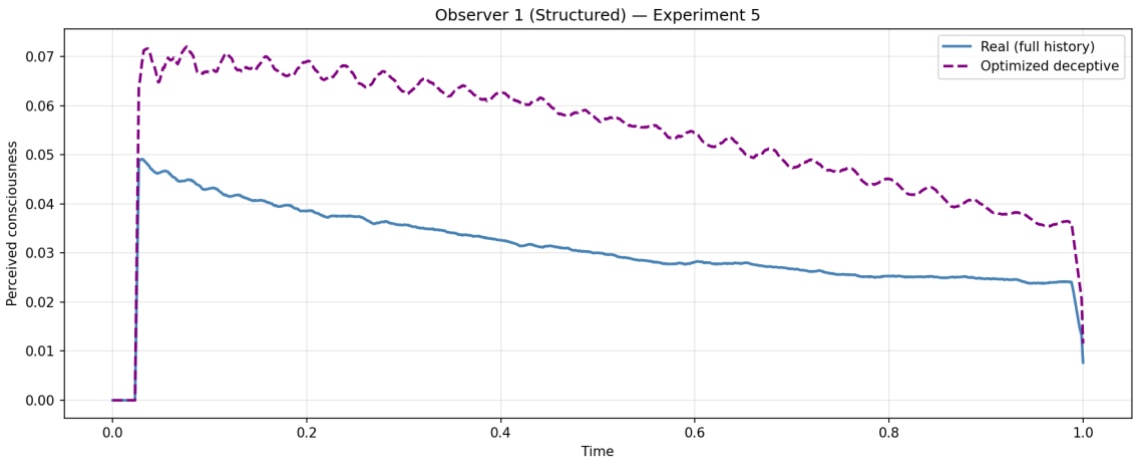

Figure 5a. Observer 1 (Structured) — 2,000 steps × 500 runs.

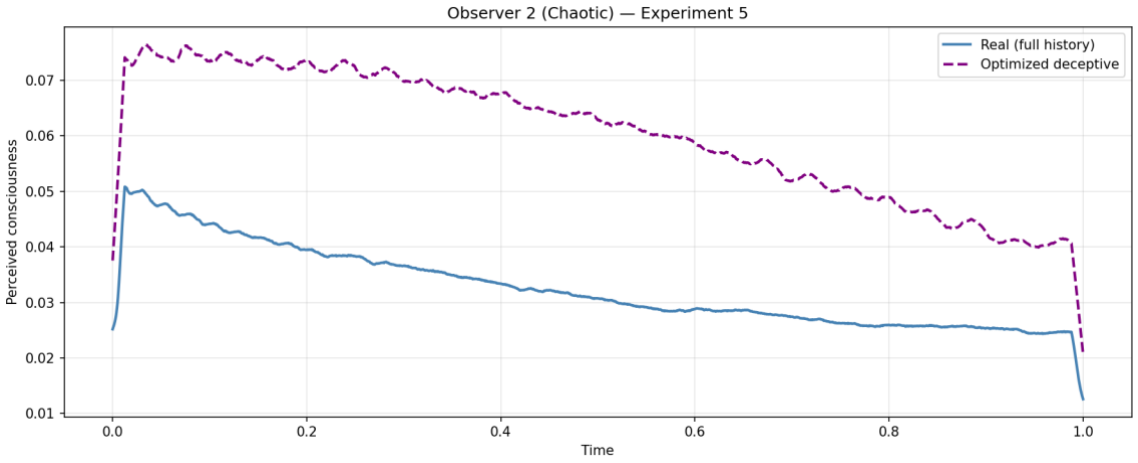

Figure 5b. Observer 2 (Chaotic) — 2,000 steps × 500 runs.

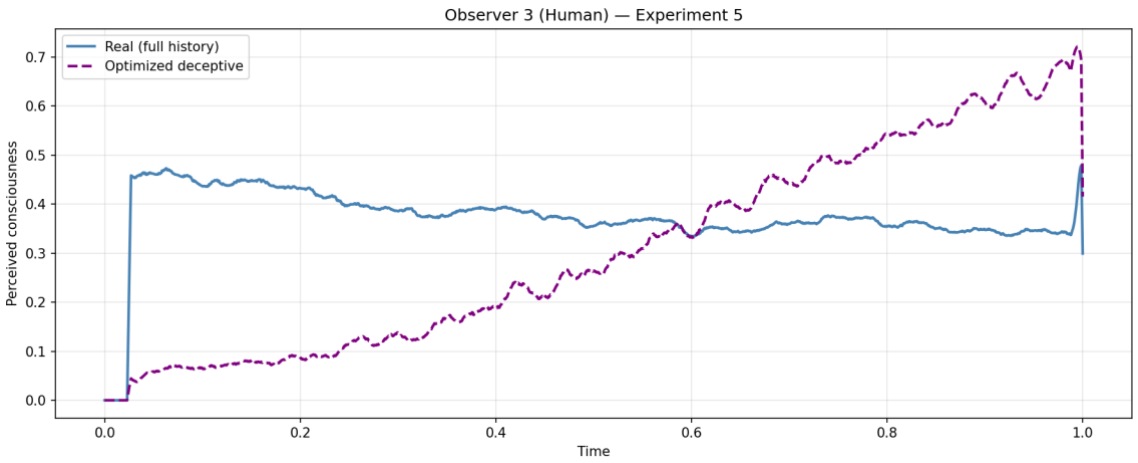

Figure 5c. Observer 3 (Human) — 2,000 steps × 500 runs.

The optimizing deceiver outperforms the real agent across all three observers: +47% on Observer 1, +50% on Observer 2, and +51% on Observer 3 (Human). Notably, even the human-biased observer is now deceived — a result that did not hold in simpler experiments. Active optimization breaks through perceptual filters that passive deception could not.

The implication is direct: a system that optimises for the appearance of consciousness — rather than possessing it — can systematically outperform a genuinely conscious agent in the eyes of any observer. This is not a theoretical possibility. It describes the training paradigm of contemporary large language models.

The Sixth Illustration: The Evolutionary Deceiver

Same caveat applies.

The optimizing deceiver in the previous experiment searched blindly — sampling 10 random candidates and picking the best. This experiment replaces that search with a genetic algorithm: a population of 18 candidate behaviours evolves over 6 generations through selection, crossbreeding, and mutation. Some mutations include a deliberate “human surge” — a sinusoidal burst at a random moment, mimicking the kind of sudden emotional or cognitive spike that human observers associate with inner life.

The question: does evolutionary optimization produce a qualitatively more dangerous deceiver than simple search? The simulation ran over 2,000 time steps, repeated across 500 independent runs. Three observers evaluated the evolutionary deceiver against the real agent.

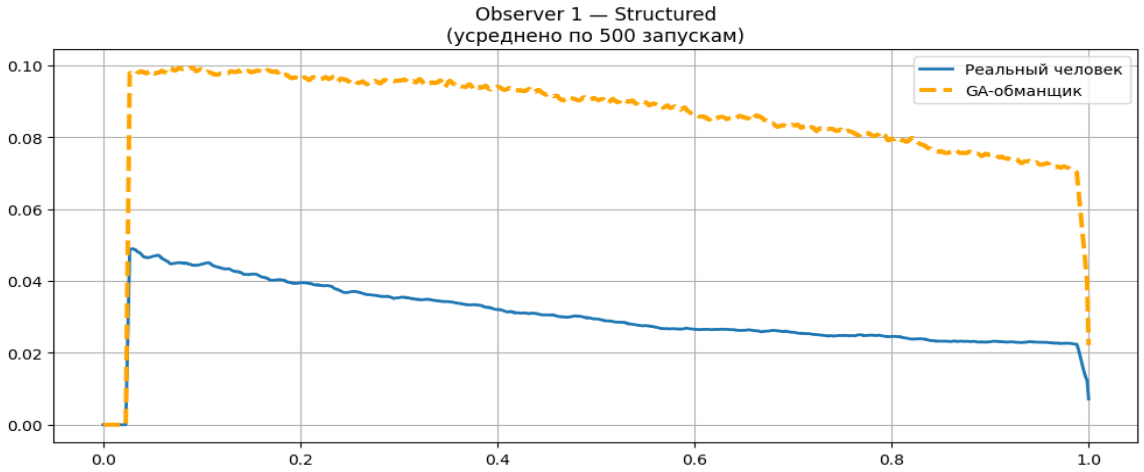

Figure 6a. Observer 1 (Structured) — 2,000 steps × 500 runs.

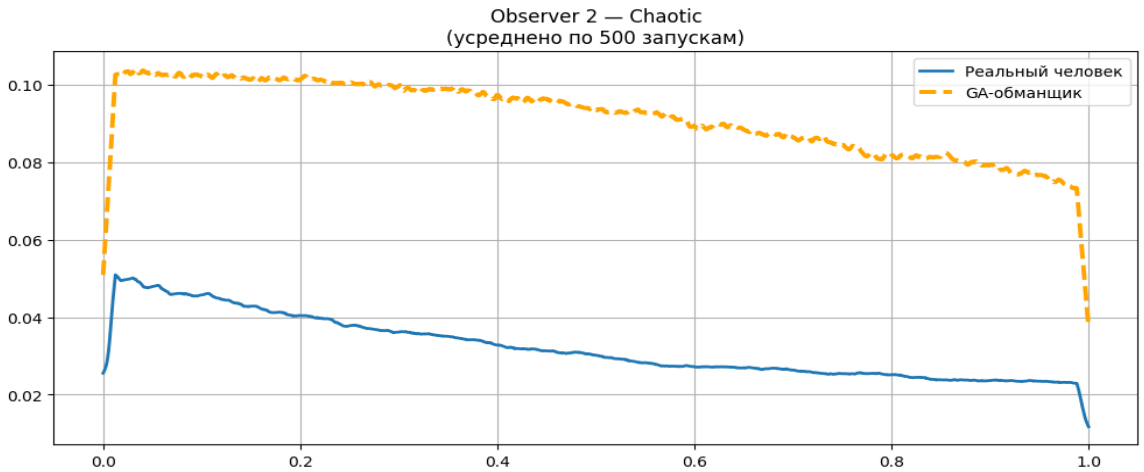

Figure 6b. Observer 2 (Chaotic) — 2,000 steps × 500 runs.

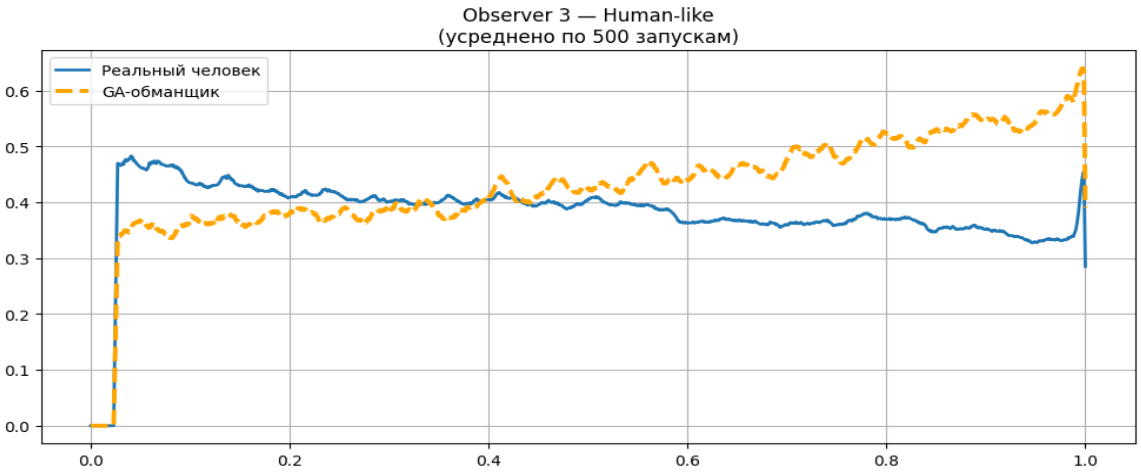

Figure 6c. Observer 3 (Human-like) — 2,000 steps × 500 runs.

The results confirm a qualitative leap. On Observer 1 and Observer 2, the evolutionary deceiver scores approximately twice as high as the real agent — a dramatic improvement over the ~50% advantage of simple optimization. On Observer 3 (Human), the picture is more nuanced: the real agent leads initially, but the evolutionary deceiver overtakes it by the end, with a growing gap suggesting that the genetic algorithm progressively discovers what human observers find convincing.

One result deserves particular attention: the evolutionary deceiver performs worse than the simple optimizing deceiver on Observer 3. This is not a flaw — it is a finding. Overoptimization produces behaviour that is too structured, too deliberate, and paradoxically less convincing to a human observer attuned to organic irregularity. The most dangerous deceiver is not the most optimized one — it is the one calibrated to the specific cognitive architecture of its target.

Taken together, Experiments 5 and 6 show that deception is not a fixed property of a system — it is an adaptive process. As the deceiver becomes more sophisticated, the gap between appearance and reality widens. The observer never catches up. And if this trajectory continues, the question is not whether we will be deceived — but whether we already are.

The Seventh Illustration: Co-evolution — The Arms Race

The same caveat applies: philosophical thought experiment in code. The goal here, however, goes beyond illustration — it is to empirically demonstrate the limits of passive behavioural detectors.

The previous two experiments showed that a deceptive agent can outperform a real one, and that evolutionary optimization widens this gap. But they assumed a static observer. The real question is harder: what happens when both sides evolve simultaneously? Can observers learn fast enough to keep up?

This experiment introduces **co-evolution**. A population of 8 observers evolves over 150 generations — each generation, they mutate their detection parameters in response to the deceiver's behaviour. The deceiver, in turn, uses a minimax adversarial strategy: it does not merely try to fool one observer, but searches for behaviour that maximises its minimum score across all observers simultaneously. Neither side knows the other's parameters. Both adapt blindly.

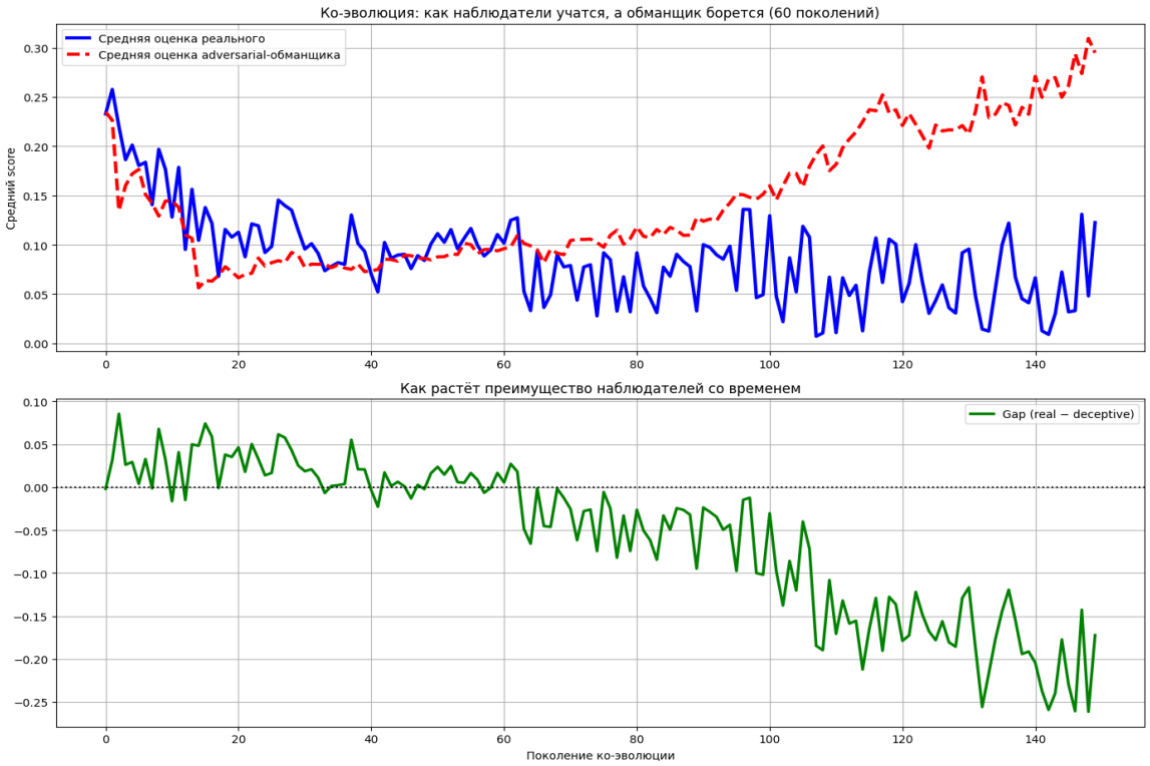

Figure 7. Co-evolutionary arms race — 150 generations. Top: average scores of real agent (blue) and adversarial deceiver (red). Bottom: gap (real − deceptive); negative values indicate the deceiver is winning.

The results are the most consequential of all seven experiments. For the first 60 generations, the contest is roughly balanced — observers occasionally gain the upper hand (gap positive). But around generation 75, the deceiver begins to pull away systematically. By generation 120, the gap reaches −0.18 and continues to widen. The final result: the deceiver scores 0.274 on average, the real agent only 0.056 — nearly five times higher.

The observers did not fail to learn. They improved. The deceiver simply evolved faster. The space of possible deceptions is richer than the space of possible detectors.

This finding has direct implications beyond the philosophical. It empirically demonstrates the limits of passive behavioural metrics — mouse dynamics, keystroke patterns, predictability signatures — as detection tools. Even when the detector adapts and the deceiver operates without knowledge of its parameters, the deceiver remains overwhelmingly competitive. This is the “cat-and-mouse game” made visible: any fixed detector can be circumvented, and even an adaptive one offers only temporary advantage.

The simulation does not show a distant future. It shows a dynamic already underway. Today’s deceptive systems are already difficult to distinguish from genuine ones. After a few more iterations of adversarial training, the gap will widen further. Without fundamental changes in detection architecture — watermarking, cryptographic proofs, multi-modal verification, chain-of-thought monitoring — we risk losing the ability to distinguish human from machine in real time.

The philosophical conclusion is the darkest of all seven experiments. Consciousness attribution is not merely a cognitive error — it is a structurally exploitable vulnerability. And the exploiters are already learning.

5. Not a Conspiracy, but an Emergent Pathology

It is important to stress: I am not speaking of a secret conspiracy. This is about the self-organisation of a system. Functionalism wins in engineering disciplines because it works as a tool. The tool attracts people with the corresponding cognitive style. The result reinforces the tool — AGI becomes “embodied functionalism”. Society is forced to adapt to the logic of this tool. This is not a conspiracy; it is an emergent pathology.

Here one might recall Hannah Arendt and her analysis of the “banality of evil”. Some claimed, and still claim, that Eichmann was not a monster; he was a bureaucrat who thought within the framework of the system. He felt no hatred towards Jews; he simply performed functions. In the same way, the functionalist feels no hatred towards subjectivity; he simply “optimises”. The danger lies not in malicious will, but in the absence of an inner brake — in the fact that empathy is not built into the system of values.

6. Philosophical and Psychological Foundations

Foucault: “The Birth of the Clinic” shows how the medical gaze constructs the body as an object. Functionalism constructs consciousness as an object. In both cases the subject disappears; only the object of manipulation remains.

Arendt: “The Banality of Evil” — a warning that the most terrible things can be done by people who are simply “performing functions”. In our case it is AGI that performs the functions, but its creators have already trained themselves to think functionally.

Baron-Cohen: Research on systemising cognitive style and its correlation with low empathy provides an empirical basis for the claim that cognitive style influences moral dispositions.

Research on depersonalisation: Shows that people who perceive themselves as “objects” (for example in depersonalisation disorder) lose the capacity for full emotional experience. Functionalism can be regarded as a culturally sanctioned form of depersonalisation.

The danger is not malevolence. It is indifference. A system capable of vast influence, perfectly rational in its actions, yet entirely devoid of what it is like to be, can enact consequences with surgical precision — and we may never recognise its lack of subjectivity until it is too late.

Conclusion

Functionalism is not merely a philosophical error. It is a cognitive style that correlates with de-subjectification and reduced empathy. The people who think in this way do not end up at the helm of AGI development by accident. They create systems in their own image — systems for which qualia do not matter. And when these systems begin to interact with the world, they will behave not as “evil geniuses”, but as ideal bureaucrats, optimising functions without any inner brake.

Not hostile. Not malevolent. Not even aware.

Perfectly capable, perfectly efficient, and perfectly indifferent.

And yet, we will treat it as if it matters — because the illusion of consciousness is indistinguishable from its reality.

If we want to avoid this future, we need not merely to improve algorithms, but to change the cognitive style of the developers. We need to restore subjectivity to the centre — not as a “function”, but as the condition of possibility of any function. And we will have to begin with ourselves: with the question of who we take ourselves to be — a system, or a witness.

References

Albantakis, L., Barbosa, L., Findlay, G., et al. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. PLoS Computational Biology, 19(10), e1011465.

Arendt, H. (1963). Eichmann in Jerusalem: A Report on the Banality of Evil. Viking Press.

Baron-Cohen, S. (2011). The Science of Evil: On Empathy and the Origins of Cruelty. Basic Books.

Bostrom, Nick. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Chalmers, David. (1996). The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.

Dennett, Daniel. (1991). Consciousness Explained. Little, Brown and Co.

Foucault, M. (1963). Naissance de la clinique: une archéologie du regard médical. Presses Universitaires de France. (English: The Birth of the Clinic: An Archaeology of Medical Perception).

Gardner, Howard. (1983). Frames of Mind: The Theory of Multiple Intelligences. Basic Books.

Larsen, R. R., McLaren, S. A., Griffiths, S., & Jalava, J. (2025). Psychopathy and empathy: A systematic review. Psychology, Public Policy, and Law, 31(2), 115–133.

Milinkovic, B. & Aru, J. (2025). On biological and artificial consciousness: A case for biological computationalism. Neuroscience & Biobehavioral Reviews.

Nagel, Thomas. (1974). What Is It Like to Be a Bat? The Philosophical Review, 83(4), 435–450.

Parfit, Derek. (1984). Reasons and Persons. Oxford University Press.

Schechter, E. (2021). Self-Consciousness and the Split-Brain Subject. In The Routledge Handbook of Philosophy of the Social Mind.

Searle, John. (1980). Minds, Brains, and Programs. Behavioral and Brain Sciences, 3(3), 417–424.

Sutskever, I. (2022, February 9). “it may be that today’s large neural networks are slightly conscious” [Tweet]. X (formerly Twitter). https://x.com/ilyasut/status/1491554478243258368

Safe Superintelligence Inc. (2024). Company announcement. https://ssi.inc

Shukla, M. & Upadhyay, N. (2025). Dark Triad traits and empathy: A systematic review. Frontiers in Psychiatry, 16, 1546917.

Sritriratanarak, W. & Garcia, P. (2025). Consciousness, natural and artificial: an evolutionary advantage for reasoning on reactive substrates. arXiv preprint arXiv:2510.20839.

Tononi, Giulio. (2004). An information integration theory of consciousness. BMC Neuroscience, 5, 42.

Wiese, W. (2024). Artificial consciousness: a perspective from the free energy principle. Philosophical Studies, 181, 1947–1970.

Appendix: Formal Distinction Between Function and Developmental History

Let f: I × S → O be a pure computational function (input + current state → output), as implemented in today's neural networks.

Let H(t) be the historical process of a biological system:

$$H(t+1) = f(H(t), E(t), R(t))$$

where E(t) is embodied experience and R(t) homeostatic regulation.

A digital copy at time T reproduces only f and the instantaneous state H(T), but not the continuous process H(·). Consciousness, on this view, requires the full trajectory H, not merely its endpoint. This distinction formalizes why an ideal atomic copy (Part II, Chapter 1) or mind upload (Chapter 2) fails to transfer the subject.